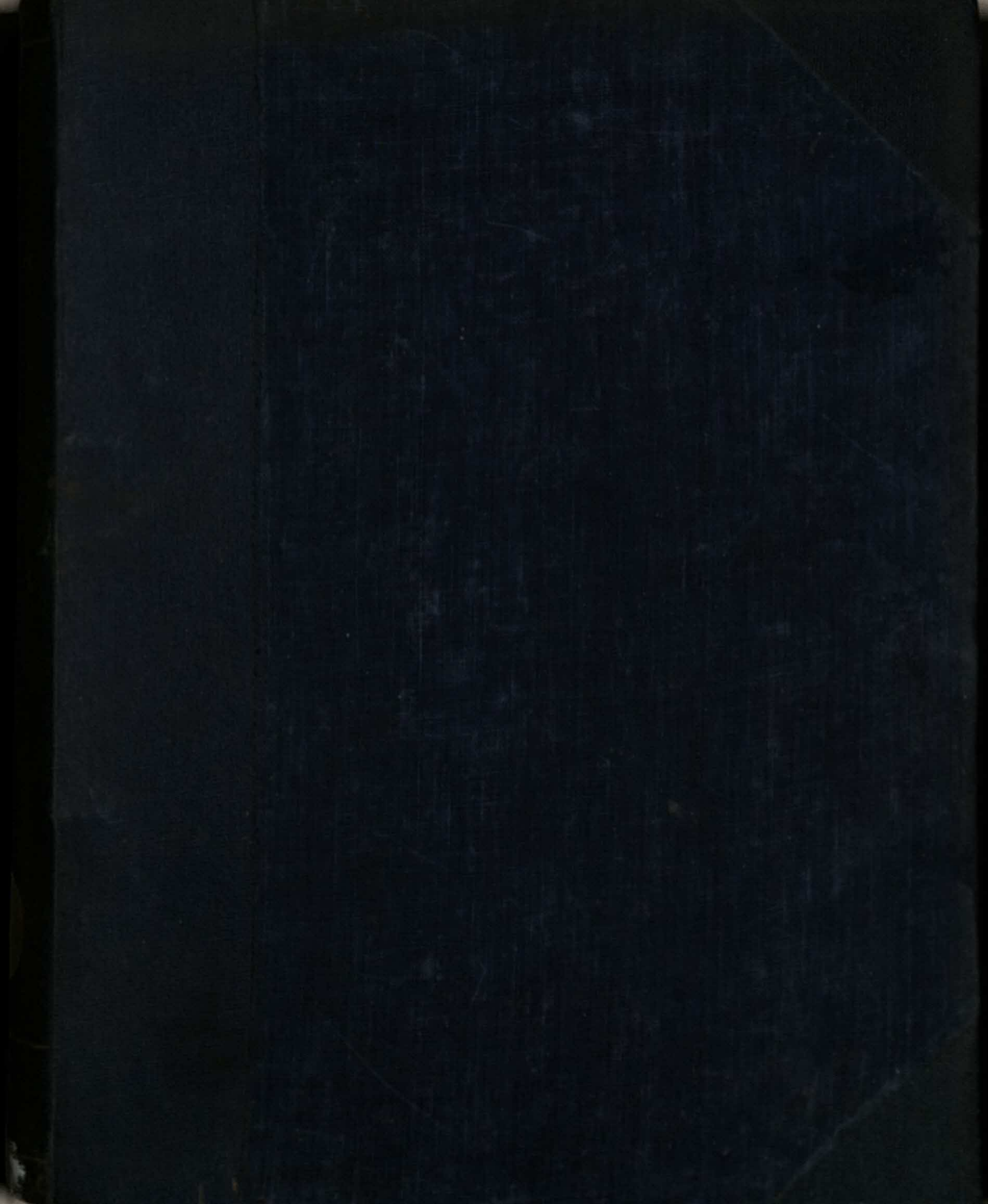




SANKHYĀ

THE INDIAN JOURNAL OF STATISTICS

SERIES A

VOLUME 23, PART ~~I~~



EDITED BY
P. C. MAHALANOBIS

FEBRUARY, 1961



STATISTICAL PUBLISHING SOCIETY
CALCUTTA



SANKHYĀ, SERIES A, VOL. 23, PART 1, 1961

CONTENTS

(All rights reserved)

	PAGE
FOREWORD	1
SAMPLING THE REFERENCE SET	3
SIMPLE APPROXIMATION TO THE PROBABILITY INTEGRAL AND $(P\chi^2, 1)$ WHEN BOTH ARE SMALL	9
MECHANISTIC MODEL OF A RANDOM PHENOMENON	11
A STUDY OF LARGE SAMPLE TEST CRITERIA THROUGH PROPERTIES OF EFFICIENT ESTIMATES	25
A METHOD OF FRACTILE GRAPHICAL ANALYSIS	41
ON SOME PROPERTIES OF ERROR AREA IN THE FRACTILE GRAPH METHOD	65
ON SOME LIMIT DISTRIBUTION CONNECTED WITH FRACTILE GRAPHICAL ANALYSIS	79
SOME LIMIT THEOREMS IN REGRESSION THEORY	91

SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS

CONTENTS OF SERIES A, VOLUME TWENTYTHREE, 1961

(All rights reserved)

PAPERS

	Page
Foreword. <i>By P. C. Mahalanobis</i>	1
Sampling the reference set. <i>By Sir Ronald A. Fisher</i>	3
Simple approximations to the probability integral and $P(\chi^2, 1)$ when both are small. <i>By J. B. S. Haldane</i>	9
Mechanistic model of a random phenomenon. <i>By M. Kalecki</i>	11
A study of large sample test criteria through properties of efficient estimates. <i>By C. Radhakrishna Rao</i>	25
A method of fractile graphical analysis. <i>By P. C. Mahalanobis</i>	41
On some properties of error area in the fractile graph method. <i>By K. Takeuchi</i>	65
Some limit distributions connected with fractile graphical analysis. <i>By J. Sethuraman</i>	79
Some limit theorems in regression theory. <i>By K. R. Parthasarathy and P. K. Bhattacharya</i>	91
The weighted mean of two normal samples with unknown variance ratio. <i>By Sir Ronald A. Fisher</i>	103
A note on mutually orthogonal Latin squares. <i>By S. S. Shrikhande</i>	115
A study of BIB designs with replications 11 to 15. <i>By C. Radhakrishna Rao</i>	117
A table of minimum functions for generating Galois fields of $GF(p^n)$. <i>By J. D. Alanen and D. E. Knuth</i>	128
Analysis of two-way designs. <i>By J. Roy and K. R. Shah</i>	129
On some problems of fractile graphical analysis. <i>By N. M. Mitrofanova</i>	145
Some remarks concerning the expectation of the error area in fractile analysis. <i>By Yukiyoji Kawada</i>	155

	Page
Minimax tests for the rate of a Poisson process and the bias rate of a normal process. <i>By J. V. Breakwell</i>	161
Minimum variance unbiased estimation in coin tossing problems. <i>By T. V. Narayana and Y. S. Sathe</i>	183
Conflicting criteria of 'goodness' of statistics. <i>By J. Sethuraman</i>	187
Sampling experiments on the combination of independent χ^2 -tests. <i>By Nikhilesh Bhattacharya</i>	191
A table of the normal integral with complex argument. <i>By Peter Ihm</i>	197
A further contribution to the numerical evaluation of certain multivariate normal integrals. <i>By Peter Ihm</i>	205
An elementary theory of the Brownian motion random function. <i>By J. Kampé De Fériet</i>	207
On testing the mean of a discrete linear process. <i>By K. R. Parthasarathy</i>	221
On a problem of Bartholomew in life testing. <i>By P. S. Swamy and S. A. D. C. Doss</i>	225
Studentization of two-stage sample means from normal populations with unknown common variance. <i>By Harold Ruben</i>	231
Testing for the single outlier in a regression model. <i>By K. S. Srikantan</i>	251
Minimal sufficient statistics for the balanced incomplete block design under an Eisenhart model II. <i>By David L. Weeks and Franklin A. Graybill</i>	261
Asymptotic bias and variance of ratio estimates in generalized power series distributions and certain applications. <i>By G. P. Patil</i>	269
Method of constructing two mutually orthogonal Latin squares of order $3n+1$. <i>By P. Kesava Menon</i>	281
Combinatorial arrangements analogous to orthogonal arrays. <i>By C. Radhakrishna Rao</i>	283
On the structure and combinatorial properties of certain semi-regular group divisible designs. <i>By W. S. Saraf.</i>	287
A factorial approach to construction of true cost of living index and its application in studies of changes in national income. <i>By K. S. Banerjee</i>	297
Generation of random permutations of given number of elements using random sampling numbers. <i>By C. Radhakrishna Rao</i>	305
The transformation of a distribution under selection. <i>By D. J. Finney</i>	309

	Page
Analysis of errors in censuses and surveys with special reference to experience in India. <i>By P. C. Mahalanobis and D. B. Lahiri</i> ...	325
On confidence interval for two-means problem based on separate estimates of variances and tabulated values of t -table. <i>By Saibal Banerjee</i> ...	359
Some limit theorems for joint distributions. <i>By J. Sethuraman</i> ...	379
Best one-sided bounds for infinitely divisible random variables. <i>By Howard G. Tucker</i> ...	387
Tchebycheff-type inequalities in terms of the mean deviation. <i>By Gerald J. Glasser</i> ...	397
A note on functional minimality of sufficient statistics. <i>By Edward W. Barankin</i> ...	401
Minimax solutions to trichotomies. <i>By L. L. Lasman</i> ...	405
Use of 'order-statistic' in sampling without replacement. <i>By P. K. Pathak</i> ...	409
On the evaluation of moments of distinct units in a sample. <i>By P. K. Pathak</i> ...	415
Moments of first order autocorrelation of the second order autoregressive scheme. <i>By Sukumar Mukherjee</i> ...	421

CORRIGENDA

Corrigenda to a paper "Tables of random normal deviates." <i>By J. M. Sengupta and Nikhilesh Bhattacharya</i> ...	308
Corrigenda to a paper "On some classification problems-I." <i>By S. John</i> ...	308
Corrigenda to a paper "On some classification statistics." <i>By S. John</i> ...	308
Corrigenda to a paper "A note on mutually orthogonal Latin squares." <i>By S. S. Shrikhande</i> ...	426

AUTHOR INDEX

<i>Alanen, J. D. and Knuth, D. E.</i> A table of minimum functions for generating Galois fields of $GF(p^n)$...	128
<i>Banerjee, K. S.</i> A factorial approach to construction of true cost of living index and its application in studies of changes in national income ...	297
<i>Banerjee, Saibal.</i> On confidence interval for two-means problem based on separate estimates of variances and tabulated values of t -table ...	359
<i>Barankin, Edward W.</i> A note on functional minimality of sufficient statistics	401

	Page
<i>Bhattacharya, P. K. and Parthasarathy, K. R.</i> Some limit theorems in regression theory	91
<i>Bhattacharya, Nikhilesh.</i> Sampling experiments on the combination of independent χ^2 -tests	191
<i>Bhattacharya, N. and Sengupta, J. M.</i> Corrigenda to a paper "Tables of random normal deviates"	308
<i>Breakwell, J. V.</i> Minimax tests for the rate of a Poisson process and the bias rate of a normal process	161
<i>Doss, S. A. D. C. and Swamy, P. S.</i> On a problem of Bartholomew in life testing	225
<i>Férier, J. Kampé De.</i> An elementary theory of the Brownian motion random function	207
<i>Finney, D. J.</i> The transformation of a distribution under selection ...	309
<i>Fisher, Sir Ronald A.</i> Sampling the reference set	3
———— The weighted mean of two normal samples with unknown variance ratio	103
<i>Glasser, Gerald J.</i> Tchebycheff-type inequalities in terms of the mean deviation	397
<i>Graybill, Franklin A. and Weeks, David L.</i> Minimal sufficient statistics for the balanced incomplete block design under an Eisenhart model II ...	261
<i>Haldane, J. B. S.</i> Simple approximations to the probability integral and $P(\chi^2, 1)$ when both are small	9
<i>Ihm, Peter.</i> A table of the normal integral with complex argument ...	197
———— A further contribution to the numerical evaluation of certain multivariate normal integrals	205
<i>John, S.</i> Corrigenda to a paper "On some classification problems I" ...	308
———— Corrigenda to a paper "On some classification statistics" ...	308
<i>Kalecki, M.</i> Mechanistic model of a random phenomenon	11
<i>Kawada, Yukiyo.</i> Some remarks concerning the expectation of the error area in fractile analysis	155
<i>Knuth, D. E. and Alanen, J. D.</i> A table of minimum functions for generating Galois fields of $GF(P^n)$	128

	Page
<i>Lahiri, D. B. and Mahalanobis, P. C.</i> Analysis of errors in censuses and surveys with special reference to experience in India ...	325
<i>Lasman, L. L.</i> Minimax solutions to trichotomies ...	405
<i>Mahalanobis, P. C.</i> Foreword ...	1
———— A method of fractile graphical analysis ...	41
<i>Mahalanobis, P. C. and Lahiri, D. B.</i> Analysis of errors in censuses and surveys with special reference to experience in India ...	325
<i>Menon, P. Kesava.</i> Method of constructing two mutually orthogonal Latin squares ...	281
<i>Mitrofanova, N. M.</i> On some problems of fractile graphical analysis ...	145
<i>Mukherjee, Sukumar.</i> Moments of first order autocorrelation of the second order autoregressive scheme ...	421
<i>Narayana, T. V. and Sathe, Y. S.</i> Minimum variance unbiased estimation in coin tossing problems ...	183
<i>Parthasarathy, K. R.</i> On testing the mean of a discrete linear process ...	221
<i>Parthasarathy, K. R. and Bhattacharya, P. K.</i> Some limit theorems in regression theory ...	91
<i>Pathak, P. K.</i> Use of 'order-statistic' in sampling without replacement ...	409
———— On the evaluation of moments of distinct units in a sample ...	415
<i>Patil, G. P.</i> Asymptotic bias and variance of ratio estimates in generalized power series distributions and certain applications ...	269
<i>Rao, C. Radhakrishna.</i> A study of large sample test criteria through properties of efficient estimates ...	25
———— A study of BIB designs with replications 11 to 15 ...	117
———— Combinatorial arrangements analogous to orthogonal arrays ...	283
———— Generation of random permutations of given number of elements using random sampling numbers ...	305
<i>Roy, J. and Shah, K. R.</i> Analysis of two-way designs ...	129
<i>Ruben, Harold.</i> Studentization of two-stage sample means from normal populations with unknown common variance ...	231
<i>Saraf, W. S.</i> On the structure of combinatorial properties of certain semi-regular group divisible designs ...	287

	Page
<i>Sathe, Y. S. and Narayana, T. V.</i> Minimum variance unbiased estimation in coin tossing problems	183
<i>Sethuraman, J.</i> Some limit distributions connected with fractile graphical analysis	79
————— Conflicting criteria of 'goodness' of statistics	187
————— Some limit theorems for joint distributions	379
<i>Sengupta, J. M. and Bhattacharya, N.</i> Corrigenda to a paper "Tables of random normal deviates"	308
<i>Shah, K. R. and Roy, J.</i> Analysis of two-way designs	129
<i>Shrikhande, S. S.</i> A note on mutually orthogonal Latin squares	115
————— Corrigenda to a paper "A note on mutually orthogonal Latin squares"	426
<i>Srikantan, K. S.</i> Testing for the single outlier in a regression model	251
<i>Swamy, P. S. and Doss, S. A. D. C.</i> On a problem of Bartholomew in life testing	225
<i>Takeuchi, K.</i> On some properties of error area in the fractile graph method	65
<i>Tucker, Howard G.</i> Best one-sided bounds for infinitely divisible random variables	387
<i>Weeks, David L. and Graybill, Franklin A.</i> Minimal sufficient statistics for the balanced incomplete block design under an Eisenhart model II	261

BIOMETRIKA

Vol. 47, Parts 3 and 4

December 1960

CONTENTS

The properties of a stochastic model for the predator-prey type of interaction between two species: *Leslie, P. H. and Gower, J. C.*; Hierarchical birth and death processes I. Theory: *Blom, Gunnar*; Hierarchical birth and death processes II. Applications: *Blom, Gunnar*; A comparison of the effectiveness of tournaments: *Glen, W. A.*; Supplemented balance: *Pearce, S. C.*; The efficiency of blocking in incomplete block designs: *Folks, John Leroy and Kempthorne, Oscar*; Queueing with balking, II: *Haight, Frank A.*; On comparing different tests of the same hypothesis: *David, H. A. and Perez, Carmen A.*; The performance of some correlation coefficients for a general bivariate distribution: *Farlie, D. J. G.*; Two expansions for the quadrivariate normal integral: *McFadden, J. A.*; On a problem connected with quadratic regression: *Laha, R. G. and Lukacs, E.*; Small sample behaviour of certain tests of the hypothesis under variance heterogeneity: *McCullough, Roger S., Gurland, John and Rosenberg, Lloyd*; Power of some two-sample non-parametric tests: *Sukhatme, Balkrishna V.*; Sampling inspection of continuous processes with no autocorrelation between successive results: *Ewan, W. D. and Kemp, K. W.*; Table of Neyman-shortest unbiased confidence intervals for the binomial parameter: *Blyth, Colin R. and Hutchinson, David W.*; On the power function of the exact tests for the 2×2 contingency table: *Bennett, B. M. and Hsu, P.*; Table of the bounds of the probability integral when the first four moments are given: *Simpson, J. A. and Welch, B. L.*; Normal approximation to the chi-square and non-central F probability functions: *Severo, Norman C. and Zelan, Marvin*; Some properties of the distributions of the logarithm of non-central F : *Barton, D. E., David, F. N. and O'Neill, Anne F.*; Tables for making inferences about the variance of a normal distribution: *Lindley, D. V., East, D. A. and Hamilton, P. A.*; Tables for the solution of the exponential equation, $\exp(-a) + ka = 1$: *Barton, D. E., David, F. N. and Merrington, M.*

The subscription, payable in advance, is 54/- (or \$8.00), per volume (including postage). Cheques should be made payable to *Biometrika*, crossed "*A/c Biometrika Trust*" and sent to the Secretary, *Biometrika Office*, Department of Statistics, University College, London, W. C. 1. All foreign cheques must be drawn on a Bank having a London agency.

Issued by THE BIOMETRIKA OFFICE, University College, London.

JOURNAL OF THE ROYAL STATISTICAL SOCIETY

SERIES B (METHODOLOGICAL), 30s each Part. Annual Subscription

£3.2s (US \$9). Post free

CONTENTS OF VOL. 22, PART 2, 1960

Series B (Methodological)

Models in the Analysis of Variance: *R. L. Plackett* (With Discussion); Discrete Stochastic Processes in Population Genetics: *W. F. Bolmer* (With Discussion); A Problem of Delayed Service—I: *A. Ben-Israel and P. Naor*; A Problem of Delayed Service—II: *A. Ben-Israel and P. Naor*; On the Transient Behaviour of a simple Queue: *P. D. Finch*; Queueing at a Single Serving Point with Group Arrival: *B. W. Conolly*; On some extensions of Bayesian Inference proposed by Mr. Lindley: *R. A. Fisher*; A Method of Estimation of Missing Values in Multivariate Data Suitable for use with an Electronic Computer: *S. F. Buck*; Estimation of Parameters of a Multivariate Normal Population from Truncated and Censored Populations: *Naunihal Singh*; Necessary Restrictions for Distributions *a posteriori*: *D. A. Sprott*; Weight of Evidence, Corroboration, Explanatory Power, Information and the Utility of Experiments: *I. J. Good*; The Logistic Process: Tables of the Stochastic Epidemic Curve and Applications: *Edwin Mansfield and Carlton Hensley*; Some Notes of Pistimetric Inference: *A. D. Roy*; Evaluation of Determinants, Characteristic Equations and their Roots for a Class of Patterned Matrices: *S. N. Roy, B. G. Greenberg and A. E. Sarhan*; Bounds for the Expected Sample Size in a Sequential Probability Ratio Test: *M. N. Ghosh*; A Note on Tests of Homogeneity Applied after Sequential Sampling: *D. R. Cox*; The Interaction Algorithm and Practical Fourier Analysis: an Addendum: *I. J. Good*; Approximate Solutions of Green's Type for Univariate Stochastic Processes: *H. E. Daniels*; The Wilcoxon Test and Non-null Hypotheses: *G. B. Wetherill*.

Obtainable through any bookseller or direct from the

ROYAL STATISTICAL SOCIETY

21 Bentinck Street, London, W. 1.

THE ANNALS OF MATHEMATICAL STATISTICS

(Founded by H. C. Carver)

The Official Journal of the Institute of Mathematical Statistics

CONTENTS OF VOLUME 31, NO. 4, DECEMBER 1960

Our Silver Anniversary: *Allen T. Craig*; Simplex-Sum Designs: A Class of Second Order Rotatable Designs Derivable from those of First Order: *G. B. P. Box and D. W. Behnken*; Third Order Rotatable Designs in Three Dimensions: *Norman R. Draper*; A Third Order Rotatable Design in Four Dimensions: *Norman R. Draper*; Some Aspects of Weighing Designs: *Damaraju Raghavarao*; Random Allocation Designs I: On General Classes of Estimation Methods: *A. P. Dempster*; A Mixed Model for the Complete Three-Way Layout with Two Random-Effects Factors: *J. P. Imhof*; Tests for Regression Coefficients when Errors are Correlated: *M. M. Siddiqui*; Mixed Model Variance Analysis with Normal Error and Possibly Non-Normal Other Random Effects, Part I: The Univariate Case: *S. N. Roy and Whitfield Cobb*; Mixed Model Variance Analysis with Normal Error and Possibly Non-Normal Other Random Effects, Part II: The Multivariate Case: *S. N. Roy and Whitfield Cobb*; An Associated Polynomial for Least Square Approximations: *Richard Warren*; Asymptotic Rate of Discrimination for Markov Processes: *Lambert H. Koopmans*; Equalities for Stationary Processes Similar to an Equality of Wald: *Shu-Teh Chen Moy*; Multivariate Chebyshev Inequalities: *Albert W. Marshall and Ingram Olkin*; On Deviations of the Sample Mean: *R. R. Bahadur and R. Ranga Rao*; On the Independence of a Sample Central Moment and the Sample Mean: *R. G. Laha, E. Lukacs and M. Newman*; Order Statistics of Partial Sums: *J. G. Wendel*; Probability Distributions Related to Random Mappings: *Bernard Harris*; Some Asymptotic Results for a Coverage Problem: *Max Halperin*; Statistical Programming: *D. F. Votaw, Jr.*; The Life Distribution and Reliability of a System with Spare Components: *Donald F. Morrison and H. A. David*; Maximizing the Probability that Adjacent Order Statistics of Samples from Several Populations form Overlapping Intervals: *Richard Cohn, Frederick Mosteller, John W. Pratt, and Maurice Tatsuoka*; Normalizing the Noncentral t and F Distributions: *Nico F. Laubscher*; Probability Content of Regions Under Spherical Normal Distributions, II: The Distribution of the Mean in Normal Samples: *Harold Ruben*; Tables of Range and Studentized Range: *H. Leon Harter*; Asymptotic Formulae for the Distribution of Hotelling's Generalized T_0^2 Statistic, II: *Koichi Ito*; On Unbiased Estimation: *L. Schmetterer*; Two-Stage Experiments for Estimating a Common Mean: *Donald Richter*; Rank-Sum Tests for Dispersions: *A. R. Ansari and R. A. Bradley*; A Relationship between Hodges' Eivariate Sign Test and a Non-Parametric Test of Daniels': *Bruce M. Hill*; Minimax Sequential Tests of Some Composite Hypotheses: *M. H. DeGroot*; Inverse Binomial Sampling Plans When an Exponential Distribution is Sampled with Censoring: *Jack Nadler*.

Subscription Rate: \$ 10.00 per year

Address orders for subscriptions and back numbers to the
Secretary, Institute of Mathematical Statistics,
Department of Statistics, University of
North Carolina, Chapel Hill, North Carolina.

ECONOMETRICA

Vol. 28, No. 4

October, 1960

CONTENTS

A Simulation of the United States Economy in Recession: *James Duesenberry, Otto Eckstein and Gary Fromm*; Liquidity Functions in the American Economy: *Martin Bronfenbrenner and Thomas Mayer*; Symposium: Simultaneous Equation Estimation: Any Verdict Yet? *Carl Christ*; Simultaneous Equations—Any Verdict Yet? *Clifford Hildreth*; Single Equation Vs. Equation System Methods of Estimation in Econometrics: *L. R. Klein*; Underidentification, Structural Estimation, and Forecasting: *Ta-Chung Liu*; Market Mechanisms and Mathematical Programming: *Hirofumi Uzawa*; On a Method of Computing the Engel Elasticities from the Concentration Curves: *N. Sreenivasa Iyengar*; A Dual Stability Theorem: *Daie W. Jorgenson*; Expansion Paths for Some Production Functions: *Earl O. Heady and John Peseke*; Notes Regarding Comments by Professor Holbrook Working on Cowles-Jones 1937 *Econometrica* Paper: *Alfred Cowles*; Note on the Correlation of First Differences of Averages in a Random Chain: *Holbrook Working*; A Note on Seasonal Inventories: *Edwin Mills*. Announcements and Notes.

Subscription price for nonmembers: \$ 12.00 per year in the United States and Canada;

\$ 10.00 per year in all other Countries. Single copies \$ 3.00

Graduate Students recommended by a faculty member: \$ 4.00 per year.

Subscriptions should be sent to

The Secretary
ECONOMETRIC SOCIETY

Box 1264, Yale Station, New Haven, Connecticut, U.S.A.

SANKHYĀ

THE INDIAN JOURNAL OF STATISTICS

Edited by : P. C. MAHALANOBIS

SERIES A, VOL. 23

FEBRUARY 1961

PART 1

FOREWORD

I have great pleasure in announcing that in future *Sankhyā* : The Indian Journal of Statistics would be published in two separate series. Series A will contain articles with emphasis on methods and techniques, and Series B on applications, data, and records.

I may recall briefly how the first number of *Sankhyā* was published. Since 1932, regular scientific meetings were being organised by Indian Statistical Institute at which statistical papers were presented. There was a pressing need for a journal in which such papers could be published; and the Council of the Indian Statistical Institute decided to start *Sankhyā* as the official organ of the Institute. The first number of *Sankhyā* was issued in June 1933; and since then the Journal is being published, by the Statistical Publishing Society, as the official organ of the Indian Statistical Institute.

The meaning of the word *Sankhyā* and the scope of the Journal, was explained as follows in the Editorial published in Volume 1 of *Sankhyā*.

“We believe that the idea underlying the integral concepts of statistics finds adequate expression in the ancient Indian word *sankhyā*. In Sanskrit the usual meaning is ‘number’, but the original root meaning was ‘determinate knowledge’.¹ In the *Atharva-Veda*, a derivative form *sankhyātā* occurs both in the sense of ‘well-known’ as well as ‘numbered’.

The history of the word *Sankhyā* shows the intimate connexion which has existed for more than 3000 years in the Indian mind between ‘adequate knowledge’ and ‘number’.²

¹ The word is derived from *khyā* (‘to perceive, view’; ‘to be known’, ‘to make well-known’ in Monier Williams’s Dictionary). The root meaning is ‘determinate knowledge’, ‘deliberation’ or ‘whatever helps us in obtaining determinate knowledge’ according as the *krt* suffix is taken in the active or instrumental form. From the latter phase is derived the technical meaning of ‘number’.

² *Atharva-Veda*, 4.25.2. It also occurs in 4.16.5 and 12.3.28. Winternitz after a full discussion of the date of the Vedic age says “we shall probably have to date the beginning of this development about 2000 or 2500 B.C., and the end of it between 750 and 500 B.C.”: *A History of Indian Literature* (English Translation, Calcutta University, 1927), Vol. I, p. 310. While the present form of the *Atharva-Veda* is believed to be later than that of the *Rg-Veda*, much of its material is considered to be as old as, if not older than, many portions of the *Rg-Veda*. (Winternitz, p. 127). The word *sankhyā* was in common use in the sense of number in the time of *Pāṇini*.

As we interpret it, the fundamental aim of statistics is to give determinate and adequate knowledge of reality with the help of numbers and numerical analysis. The word *Sankhyā* embodies the same idea, and this is why we have chosen this name for the Indian Journal of Statistics. The spirit and outlook of *Sankhyā* will be universal, but its form and content must necessarily be, to some extent, regional.

We shall keep the special needs of India in view without, however, restricting the scope of the Journal in any way. We shall naturally devote closer attention to the collection and analysis of data relating to India, but we shall try to study all Indian questions in relation to world problems.

A research journal serves that narrow border land which separates the known from the unknown, and it is not always possible to see clearly the lines of future developments. We shall, therefore, invite papers of all kinds appraising them only on the basis of observational accuracy and logical reasoning. We shall publish carefully collected statistical materials irrespective of the subject even if they have not received any analytic treatment. We shall pay special attention to developments of the mathematical theory of statistics, and include abstracts and expositions of important papers published elsewhere. We shall try to help statistical researches on co-operative lines by bringing workers in different parts of India in contact, and by providing a medium for exchange of ideas. Bibliographies of Indian Statistical publications, numerical tables tending to reduce the labour of computation, book reviews, and notes and comments on current topics are some of the ways in which we shall try to make *Sankhyā* useful to statistical workers in India. Knowing that our resources are small we shall seek guidance and help from other countries, and we shall welcome and thankfully receive papers from abroad."

We have not been able to do all that we had in view. We are glad, however, that an increasing volume of matter is being received in recent years. The Council of the Indian Statistical Institute therefore decided that *Sankhyā* should be published from 1960 in two separate series. We take this opportunity of reminding ourselves of our original programme.

We thank the many readers and the subscribers to *Sankhyā* for their kind support and encouragement during the last twenty-seven years without which *Sankhyā* would not have occupied the position it holds today.

P. C. Mahalanobis
Editor

SAMPLING THE REFERENCE SET

By SIR RONALD A. FISHER

University of Adelaide

1. THE REFERENCE SET OF A PROBABILITY STATEMENT

Every test of significance involves probability statements, conditioned on the truth of the hypothesis. Conceptually, these can be verified by sampling the reference set which is their mathematical basis. The misapprehension has, however, been widely promulgated that such sampling involves no more than a repetition of the process by which the data to be tested came into existence. Easy examples, (Fisher, 1956-59) such as the test of significance of a linear regression, or the test of proportionality in a two by two table, have frequently been cited to show that such a method of "repeated sampling from the same population" is erroneous, and irrelevant to the test of significance which it is proposed to verify. The recognition of the appropriate reference set is an essential first step to understanding a test of significance, and, therefore, to setting up an appropriate process of possible verification. We may exemplify such a process using the long-disputed test of significance for the difference between the means of two hypothetical populations, the variances of which are not in any known ratio, when a random sample of each is available.

2. INFERENCES FROM TWO NORMAL SAMPLES

The means of the two samples will be represented by $\bar{x}_1$ and $\bar{x}_2$, where

$$(n_1+1)\bar{x}_1 = S(x_1)$$

$$(n_2+1)\bar{x}_2 = S(x_2)$$

the estimated variances of the means by s_1^2 and s_2^2 , where

$$n_1(n_1+1)s_1^2 = S(x_1 - \bar{x}_1)^2 = S_1,$$

$$n_2(n_2+1)s_2^2 = S(x_2 - \bar{x}_2)^2 = S_2,$$

these estimates being based on n_1 and n_2 degrees of freedom respectively.

The test of significance must involve as parameters the known values n_1, n_2 and s_1/s_2 ; these are, therefore, known characteristics of the reference set. If a more comprehensive set were first considered, the facts that s_1/s_2 is known to the observer, and is not irrelevant to the probability statements to be made, and that the set in which s_1/s_2 is constant contains no further relevant sub-set, may be taken as defining the set which is to be sampled. If we are to erect a sampling process capable of testing any one of the values tabulated by Behrens or Sukhatme, or others, for Behrens' test, the first step is to obtain random samples having the correct values for this ratio.

3. VERIFICATION OF BEHRENS' TEST

Now if σ_1^2 and σ_2^2 are the true variances of the populations sampled, the distribution of

$$2z = \log (n_1+1)s_1^2\sigma_2^2/(n_2+1)s_2^2\sigma_1^2$$

is known in terms of n_1 and n_2 . Let z_p stand for that value for which

$$\Pr(z < z_p) = P,$$

and let us take for P a representative series of fractions, such as

$$P_i = (2i-1)/20,000$$

where i takes integral values from 1 to 10,000.

For each of these 10,000 values

$$\sigma_2^2/\sigma_1^2 = (n_2+1)s_2^2 e^{2z}/(n_1+1)s_1^2,$$

and is, therefore, known.

We may now take two Normal populations having equal means, and variances in the ratio required.

Successive samples of the same sizes as those constituting the data may now be taken, but all will be rejected in which the experimental ratio of s_1 to s_2 does not agree, within a specified tolerance, with that originally observed. For each value of i the first which satisfies this condition may be taken as a random representative of the appropriate reference set.

If d_p is the tabular value to be verified, such random pairs of samples may be classified as satisfying one of the three inequalities,

$$(\bar{x}_1 - \bar{x}_2) < -d_p \sqrt{s_1^2 + s_2^2},$$

$$-d_p \sqrt{s_1^2 + s_2^2} < (\bar{x}_1 - \bar{x}_2) < d_p \sqrt{s_1^2 + s_2^2},$$

$$d_p \sqrt{s_1^2 + s_2^2} < (\bar{x}_1 - \bar{x}_2),$$

in which the values inserted for $\bar{x}_1$, $\bar{x}_2$, s_1^2 and s_2^2 are derived from the experimental sampling. For tabular values of the 5% point, the expected numbers in these three classes are

$$250, 9500, 250,$$

while, for values tabulated at the 1% point, the expectations are

$$50, 9900, 50$$

respectively.

SAMPLING THE REFERENCE SET

As in other cases the verification need not be carried out; it is sufficient that it is a precisely defined procedure which could be carried out with any degree of precision, and with calculable consequences.

4. THE WEIGHTED MEAN

The method set out above of generating a sample of the only possible reference set appropriate to the data may be used to illustrate the solution of some other related problems. For these we should suppose that a sample of a million pairs of samples have been obtained having the correct ratio s_1/s_2 . These pairs will differ from each other in the relative difference $d \left(= (\bar{x}_1 - \bar{x}_2) / \sqrt{s_1^2 + s_2^2} \right)$ observed, but, as in the case of the ratio s_1/s_2 , a selection can be made agreeing sufficiently with the original observations in this particular. This doubly screened sample will still show variation in the position of the true common mean relative to the two means observed, and will supply the appropriate verification of the distribution of the true mean, which as L. C. Payne has observed, depends on the four parameters $n_1, n_2, s_1/s_2$ and d and would, therefore, be exceedingly troublesome to tabulate, though capable of an effective asymptotic representation.

For this, from the primary equations

$$\mu - \bar{x}_1 = s_1 t_1,$$

$$\mu - \bar{x}_2 = s_2 t_2,$$

we may derive

$$\mu - \bar{x} = S u,$$

where

$$\bar{x} = \frac{s_2^2 \bar{x}_1 + s_1^2 \bar{x}_2}{s_1^2 + s_2^2}, \quad S = \frac{s_1 s_2}{\sqrt{s_1^2 + s_2^2}}$$

so that

$$\frac{1}{S^2} = \frac{1}{s_1^2} + \frac{1}{s_2^2},$$

and u_P is defined by the relation

$$\Pr(u < u_P) = P.$$

Then if

$$P = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{1}{2}v^2} dv,$$

u_P may be evaluated by an expansion

$$u_P = x + \sum_{r,s} u_{rs} n_1^{-r} n_2^{-s},$$

where u_{rs} is a polynomial in x , d and $\cos \theta$, $\sin \theta$, such as

$$\frac{u_{10}}{n_1} = \frac{1}{4n_1} \{ (x^3+x)c^4 + 4d(x^2+1)c^3s + (6d^2-2)xc^2s^2 + 4(d^3-d)cs^3 \}$$

$$\frac{u_{01}}{n_2} = \frac{1}{4n_2} \{ -4(d^3-d)c^3s + (6d^2-2)xc^2s^2 - 4d(x^2+1)cs^3 + (x^3+x)s^4 \}$$

in which c and s stand for the cosine and sine of θ , and u_{rs} may be obtained from its conjugate u_{sr} by interchanging c and s and reversing the sign of d .

The expansion is similar to, though owing to the additional parameter, somewhat more complex than the expansion for d (Table 2 of my 1941 paper). Unlike the distribution of d , that of the weighted mean is not symmetrical.

5. THE UNKNOWN VARIANCES

A third problem has been regarded as quite insoluble, namely to determine the simultaneous distribution of σ_1 and σ_2 , when, with such data, it is also given that the true means are equal. It may be shown without difficulty that, while the three statistics, s_1 , s_2 , d are jointly exhaustive, yet no two functions of these three provide simultaneous exhaustive estimates of σ_1 and σ_2 . Nor has any function of these quantities a sampling distribution independent simultaneously of σ_1 and σ_2 , as would be the case of an Ancillary Statistic.

Now if σ'_1 and σ'_2 stand for the standard deviations chosen for the two populations to be sampled experimentally, and s'_1 and s'_2 for the estimates of the standard deviations obtained for the means, because these have been selected so that,

$$\frac{s'_1}{s_1} = \frac{s'_2}{s_2}$$

where s_1 and s_2 are the estimates from the original data, we may obtain σ_1 and σ_2 so that

$$\frac{\sigma'_1}{\sigma_1} \text{ and } \frac{\sigma'_2}{\sigma_2}$$

are also in the same ratio.

SAMPLING THE REFERENCE SET

Within the selection that has been made to give the correct values of the statistics s_1 , s_2 and d , we now have a distribution of pairs of values of σ_1 and σ_2 , so that a statement of the form

$$\Pr(\sigma_1 < \alpha_1, \sigma_2 < \beta) = F(\alpha, \beta, s_1, s_2, d)$$

can be established.

The function on the right can be expressed as the ratio

$$\int_0^{\alpha} d\sigma_1 \int_0^{\beta} d\sigma_2 \cdot W \div \int_0^{\infty} d\sigma_1 \int_0^{\infty} d\sigma_2 \cdot W,$$

where W takes the form

$$\frac{1}{\sigma_1^{n_1+1} \sigma_2^{n_2+1} \sqrt{\sigma_1^2 + \sigma_2^2}} \exp -\frac{1}{2} \left\{ \frac{S_1}{\sigma_1^2} + \frac{S_2}{\sigma_2^2} + \frac{(\bar{x}_1 - \bar{x}_2)^2}{\sigma_1^2 + \sigma_2^2} \right\}.$$

The three statistics in the index of the exponential term are jointly exhaustive. Hence there is no other observable which could be used to define a recognizable and relevant sub-set. Without the third statistic the observations would merely justify the distribution appropriate to the equivalence,

$$\frac{S_1}{\sigma_1^2} = \chi_{n_1}^2, \frac{S_2}{\sigma_2^2} = \chi_{n_2}^2,$$

for two independent random variables χ^2 .

The additional observation

$$\bar{x}_1 - \bar{x}_2$$

introduces the factor

$$\frac{1}{\sqrt{\sigma_1^2 + \sigma_2^2}} \exp -\frac{1}{2} \left\{ \frac{(\bar{x}_1 - \bar{x}_2)^2}{\sigma_1^2 + \sigma_2^2} \right\}.$$

For variation of $\sigma_1^2 + \sigma_2^2$ this is stationary when

$$\frac{-1}{\sigma_1^2 + \sigma_2^2} + \frac{(\bar{x}_1 - \bar{x}_2)^2}{(\sigma_1^2 + \sigma_2^2)^2} = 0.$$

The factor increases with $\sigma_1^2 + \sigma_2^2$ up to a maximum when

$$\sigma_1^2 + \sigma_2^2 = (\bar{x}_1 - \bar{x}_2)^2,$$

thereafter decreasing.

The example is interesting as showing that the possibility of making simultaneous probability statements about parameters is not limited to cases of simultaneous exhaustive estimation, in the strict sense, though a jointly exhaustive set of statistics is still utilized.

REFERENCES

- FISHER, R. A. (1935): The fiducial argument in statistical inference. *Ann. Eug.*, **6**, 391-398.
- (1941): The asymptotic approach to Behrens' integral with further tables for the d test of significance. *Ann. Eug.*, **2**, 142-172.
- (1956-59): *Statistical Methods and Scientific Inference*, Oliver and Boyd, Edinburgh.

Paper received : August, 1960.

SIMPLE APPROXIMATIONS TO THE PROBABILITY INTEGRAL AND $P(\chi^2, 1)$ WHEN BOTH ARE SMALL

By J. B. S. HALDANE
Indian Statistical Institute

The most frequently used test of significance involves either the "tail" of the probability integral, or what is equivalent, the probability of χ^2 with one degree of freedom. I believe, moreover, that it is more efficient, when using tests of significance, to think in terms of $-\log Q$ (in the Pearson-Hartley 1954) notation, rather than Q . If $-\log Q < 1.3$ we do not usually regard a deviation as significant. If $-\log Q > 3$ we regard it as highly significant. Good (1950) has given an approximation to $-\log Q$, for a reduced normal variate, namely

$$-10 \log Q = 2 \frac{1}{6} x^2 + 4 + 10 \log_{10} x. \quad \dots (1)$$

which is adequate for most purposes. My own is somewhat more accurate. It is well known that for a reduced normal distribution

$$Q = \int_x^\infty (2\pi e^{t^2})^{-\frac{1}{2}} dt \\ = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2} x^{-1} (1 - x^{-2} + 1.3x^{-4} - 1.3.5x^{-6} + 1.3.5.7x^{-8} -),$$

the series being of course an asymptotic expansion which is of little value for computation till $x > 4$. Hence

$$Q = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2} (x^2 + 2 - 3x^{-2} + 16x^{-4} - 124x^{-6} + 1224x^{-8} + \dots)^{-\frac{1}{2}} \\ \ln Q = -\frac{1}{2}x^2 - \frac{1}{2}\ln(x^2 + 2) - \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln(1 - 3x^{-4} + 22x^{-6} - 168x^{-8} + 1560x^{-10} -) \\ = -\frac{1}{2}x^2 - \frac{1}{2}\ln(x^2 + 2) - \frac{1}{2}\ln 2\pi + \frac{1}{2}(3x^{-4} - 22x^{-6} + \frac{345x^{-8}}{2} - 1626x^{-10} +) \\ -\log_{10} Q = \frac{1.52x^2}{7} + \frac{1}{2} \log_{10}(x^2 + 2) + 0.399 \text{ nearly.} \quad \dots (2)$$

I use the approximation $\frac{3.04}{7} = .434286$ for $\log_{10} e = .434294$. Good's approximation is equivalent to .43. The following table shows the accuracy of (2).

x	2	2.5	3	3.5
$-\log Q$ from (2)	1.657	2.214	2.874	3.636
$-\log Q$ (correct)	1.643	2.207	2.870	3.633

It is clear that for values of x exceeding 2 the approximation is quite satisfactory, and there is no objection to using .40 instead of .399. Curiously enough the expression

$$\frac{1.52x^2}{7} + \frac{1}{2} \log(x^2 + 2 - 3x^{-2}) + .399$$

does not give any greater accuracy in the neighbourhood of 2. When $x = 2$, it gives $-\log Q = 1.628$, whose error happens to be the same as that of (2).

For χ_1^2 we have

$$Q(\chi^2|1) = \frac{1.52\chi^2}{7} + \frac{1}{2} \log_{10}(\chi^2+2) + .098 \quad \text{nearly.} \quad \dots (3)$$

We may use 0.1 for .098 with little disadvantage. Clearly the error is the same as that of (2).

A more interesting, but probably less useful, approximation to the asymptotic expansion is obtained as follows.

$$\begin{aligned} 1-x^{-2}+3x^{-4}-15x^{-6}+105x^{-8}-945x^{-10}+10395x^{-12}-135135x^{-14} \\ = 1-(x^2+3)^{-1}-6x^{-6}(1-13x^{-2}+144x^{-4}-1692x^{-6}+22401x^{-8}-) \\ = 1-(x^2+3)^{-1}-6(x^6+13x^4+25x^2+145)^{-1}-12720x^{-14}+\dots \end{aligned}$$

Unfortunately the next polynomial, if the process is repeated, does not have integral coefficients. However the expression

$$\int_x^\infty (2\pi e^{t^2})^{-\frac{1}{2}} dt = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2} x^{-1} [1-(x^2+3)^{-1}-6(x^6+13x^4+25x^2+145)^{-1}+O(x^{-14})] \quad \dots (4)$$

is more accurate than (2) for $x \geq 2$. When $x = 2$ it gives $-\log Q = 1.6411$, or $Q = .022826$, the correct value being .022750. For values of x exceeding 3 it is very accurate, and might perhaps have been used instead of the continued fraction which Sheppard (1939) actually employed. Its greatest interest is perhaps that it can be employed for the approximate summation of a number of asymptotic expansions, such as that for the "tail" of the exponential integral. The general expression is

$$\begin{aligned} 1-ht^{-1}+h(h+1)t^{-2}-h(h+1)(h+2)t^{-3}+\dots \\ = 1-h(t+h+1)^{-1}-h(h+1)[t^3+(3h+5)t^2+(h+2)(3h+1)t+(h+2)(h^2+7)]^{-1}+O(t^{-7}). \end{aligned} \quad \dots (5)$$

On putting $h = \frac{1}{2}$, $t = \frac{1}{2}x^2$, we readily find (4).

There are sound theoretical grounds for using $\log \left(\frac{Q}{1-Q} \right)$ as a measure of the credibility of a hypothesis. It is however hard to find a simple approximation to this quantity. I suggest that Q might be expressed in C.W. Allen's (1955) expression dex , ($n dex = 10^n$). Q^{-1} might be called the improbability. Thus if on a given hypothesis $P(\chi^2)$ were .02 for a particular sample, we might say that on the basis of this sample the improbability of the hypothesis was 1.7 dex .

REFERENCES

- ALLEN C. W. (1955): *Astrophysical Quantities*, Athlone Press, London.
 GOOD, L. J. (1950): *Probability and the Weighing of Evidence*, Griffin, London.
 PEARSON, E. S., and HARTLEY, H. O. (1954): *Biometrika Tables for Statisticians*, Vol. 1. Cambridge University Press.
 SHEPPARD W. F. (1939): The probability integral. *Brit. Ass. Math. Tab.*, 7, Cambridge University Press.

Paper received : April 1960.

MECHANISTIC MODEL OF A RANDOM PHENOMENON*

By M. KALECKI

Warsaw

SUMMARY. Let m denote the number of watches whose faces have a circumference equal to unity and are divided into two segments, a black and a white one. Each of the watches has only one hand, whose point moves with the velocity $\theta_1, \theta_2, \dots, \theta_m$. We check at intervals of a unit of time all the m watches to see whether their hands are in the black or in the white field. If the number of black fields is even, we record A ; if it is odd, we record B .

Let $(A)_{Nm}$ denote the frequency of the occurrence of A in N observations, and $(A_j A_{j+a})_{Nm}$ the frequency of the occurrence of the combination AA in the j -th and the $(j+a)$ -th observations. In a similar way we define the symbols $(B)_{Nm}$ and $(B_j B_{j+a})_{Nm}$, $(A_j B_{j+a})_{Nm}$ and $(B_j A_{j+a})_{Nm}$.

Under certain assumptions regarding the velocities $\theta_1, \theta_2, \dots, \theta_m$ and under the assumption that the length p_k of the arc of the black segment on the k -th watch satisfies the condition $|1 - 2p_k| \leq a < 1$ where a is constant, the author obtains the following theorems:

$$\begin{aligned} (1) \quad & \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (A)_{Nm} = \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (B)_{Nm} = \frac{1}{2}; \\ (2) \quad & \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (A_j A_{j+a})_{Nm} = \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (B_j B_{j+a})_{Nm} \\ & = \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (A_j B_{j+a})_{Nm} = \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (B_j A_{j+a})_{Nm} = \frac{1}{2}. \end{aligned}$$

which means that the j -th and the $(j+a)$ -th observations are asymptotically independent with any degree of approximation if m is sufficiently large.

(3) The j -th observation is, in the same manner, asymptotically independent of all the a observations preceding it.

I. THE PRINCIPLES OF THE "WATCH GAME"

Let us imagine m watches whose faces have a circumference equal to unity and are divided into two sectors—a black and a white. Each watch has only one hand, whose point moves with the constant velocity $\theta_1, \theta_2, \dots, \theta_m$. We check at intervals of a time unit all the m watches in order to find whether the hand is in the black or the white sector. In case the number of “black sectors” is even we record A ; in case it is odd we record B . This is one lot. The game consists of guessing whether A or B will occur. We assume, that the velocities $\theta_1, \theta_2, \dots, \theta_m$ are irrational and arithmetically independent (i.e. no θ is a linear function with rational coefficients of other θ 's).

If the initial positions of the points of the hands are recorded as zero, then the positions in subsequent N lots will be:

$$\begin{array}{ccccccc} \theta_1 - [\theta_1], & 2\theta_1 - [2\theta_1], & \dots, & N\theta_1 - [N\theta_1], \\ \theta_2 - [\theta_2], & 2\theta_2 - [2\theta_2], & \dots, & N\theta_2 - [N\theta_2], \\ \vdots & \vdots & \vdots & \vdots \\ \theta_m - [\theta_m], & 2\theta_m - [2\theta_m], & \dots, & N\theta_m - [N\theta_m]. \end{array}$$

where $[\theta]$ denotes the largest integer lesser than θ .

It should be noticed that simultaneously with the game based on the set of m watches analogous games based on the subset of the first k watches can be played.

* A Polish version of this paper appeared in *Zastosowania Matematyki*, IV, 2 (1958).

2. FREQUENCY OF OCCURRENCE OF A AND B

Notations. Let us denote the frequency of occurrence of the white sector on the k -th watch by w_{Nk} and of the occurrence of the black sector by b_{Nk} ; moreover, let us denote the length of the arc of the white sector on the k -th watch by p_k ; thus the length of the arc of the black sector will be $1-p_k$; finally, let us denote by $(A)_{Nk}$ and $(B)_{Nk}$ the frequency of the occurrence of A and B respectively in the game based on the subset of the first k watches. Thus the frequency of the occurrence of A and B in the game based on all m watches will be denoted by $(A)_{Nm}$ and $(B)_{Nm}$ respectively.

Theorem 1 : If $|1-2p_k| \leq \alpha < 1$, where α is a constant number and $k = 1, 2, \dots, m$ then $\lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (A)_{Nm} = \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (B)_{Nm} = \frac{1}{2}$.

Proof : From the theorems in the Appendix the following characteristics of the positions of hands may be derived. If, as assumed, the velocities $\theta_1, \theta_2, \dots, \theta_m$ are irrational and none of them is a linear function with rational coefficients of other θ 's we have :

(1) According to Theorem I of the Appendix

$$\lim_{N \rightarrow \infty} w_{Nk} = p_k, \quad \lim_{N \rightarrow \infty} b_{Nk} = 1 - p_k;$$

(2) According to Theorem II of the Appendix the frequency of the occurrence in one lot of the white sector on watch k and of the white sector on watch l approaches $p_k p_l$, when $N \rightarrow \infty$. The limits of frequency for the white-black, black-white and black-black combinations will be $p_k(1-p_l)$, $(1-p_k)p_l$ and $(1-p_k)(1-p_l)$, respectively. An analogous theorem will hold good for the combination including any subset of watches or all the m watches. In the terminology of the probability calculus we may say that the position of the hand of the k -th watch is asymptotically independent of the positions of hands of other watches.

Let us take into consideration the game based on the subset of the first k watches. It follows from the characteristics (1) and (2), that $(A)_{Nk}$ approaches a definite limit $(A)_k$ while $(B)_{Nk}$ approaches $(B)_k$ when $N \rightarrow \infty$.

Let us add to the subset the $(k+1)$ -th watch. In the game based on this new subset in a given lot A will occur either when in the game based on the subset of the first k watches A occurred in that lot and on the $(k+1)$ -th watch the hand stopped in the white sector; or when in the game based on the subset of the first k watches B occurred in that lot and on the $(k+1)$ -th watch the hand stopped in the black sector.

Taking into consideration the "independence" of the positions of the hands we obtain for the limit of frequency $(A)_{k+1}$

$$(A)_{k+1} = (A)_k(1-p_{k+1}) + (B)_k p_{k+1}.$$

Analogously

$$(B)_{k+1} = (B)_k(1-p_{k+1}) + (A)_k p_{k+1}.$$

As a result

$$(A)_{k+1} - (B)_{k+1} = ((A)_k - (B)_k)(1-2p_{k+1}).$$

MECHANISTIC MODEL OF A RANDOM PHENOMENON

By means of full induction we obtain :

$$(A)_m - (B)_m = (1-2p_1)(1-2p_2)\dots(1-2p_m).$$

From the assumption $|1-2p_k| \leq \alpha < 1$ it follows, that

$$\lim_{m \rightarrow \infty} ((A)_m - (B)_m) = 0.$$

Thus we have

$$\lim_{m \rightarrow \infty} (A)_m = \lim_{m \rightarrow \infty} (B)_m = \frac{1}{2}$$

or

$$\lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (A)_{Nm} = \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (B)_{Nm} = \frac{1}{2},$$

and this at any values of $p_1, p_2, \dots, p_m$, if only the condition $|1-2p_k| \leq \alpha < 1$, is fulfilled.

3. "INDEPENDENCE" OF TWO SUBSEQUENT LOTS

Let us denote the frequency of the change from black to white, or *vice versa*, in two subsequent lots (i.e. the occurrence of a white-black or black-white combination) on the k -th watch by δ_{Nk} . The frequency of "no change" in two subsequent lots (i.e. the occurrence of a white-white or black-black combination) will thus be equal to $1-\delta_{Nk}$. Moreover, let us denote the frequencies of the occurrence of the combinations AA , BB , AB and BA in two subsequent lots of the game based on the subset of the first k watches by $(AA)_{Nk}$, $(BB)_{Nk}$, $(AB)_{Nk}$, and $(BA)_{Nk}$ respectively. Accordingly the frequencies of the occurrence of AA , BB , AB and BA in two subsequent lots of the game based on all m watches will be denoted by $(AA)_{Nm}$, $(BB)_{Nm}$, $(AB)_{Nm}$ and $(BA)_{Nm}$. Finally, let us denote by Z_{Nk} the frequency of change from A to B or *vice versa* in two subsequent lots of the game based on the first k watches; thus we have :

$$Z_{Nk} = (AB)_{Nk} + (BA)_{Nk}, \quad 1 - Z_{Nk} = (AA)_{Nk} + (BB)_{Nk}.$$

For the game based on all m watches we have obviously

$$Z_{Nm} = (AB)_{Nm} + (BA)_{Nm}, \quad 1 - Z_{Nm} = (AA)_{Nm} + (BB)_{Nm}.$$

Finally, let us denote $\theta_k - [\theta_k]$ by θ'_k .

Theorem 2 : If $|1-2p_k| \leq \alpha < 1$ and $\theta'_k \geq \beta > 0$, $1-\theta'_k \geq \beta > 0$, $|\frac{1}{2}-\theta'_k| \geq \beta > 0$ where α and β are constant numbers, then

$$\lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (AA)_{Nm} = \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (BB)_{Nm} = \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (AB)_{Nm} = \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (BA)_{Nm}.$$

Proof: From the theorems in the Appendix the following characteristics of the positions of the hands may be derived :

(3) According to Theorem III in the Appendix δ_{Nk} approaches certain limit δ_K , when $N \rightarrow \infty$, or

$$\lim_{N \rightarrow \infty} \delta_{Nk} = \delta_k \quad \text{and} \quad \lim_{N \rightarrow \infty} (1-\delta_{Nk}) = 1-\delta_k.$$

At the same time according to Theorem V in the Appendix, if $|1-2p_k| \leq \alpha < 1$ and $\theta'_k \geq \beta > 0$, $1-\theta'_k \geq \beta > 0$, $|\frac{1}{2}-\theta'_k| \geq \beta > 0$, where α and β are constant numbers then $|1-2\delta_k| \leq \lambda < 1$, where λ is also a constant number.

(4) According to Theorem IV of the Appendix the frequency of the occurrence of "change" (or "no change") in two subsequent positions of the hand of the k -th watch is "independent" of the frequency of the "change" (or "no change") on other watches. For instance the limit of the frequency combination of "change" on the k -th watch and of "no change" on the l -th watch amounts to $\delta_k(1-\delta_l)$.

Let us now take into consideration the game based on the subset of the first k watches. It follows from the characteristics (3) and (4) that if $N \rightarrow \infty$, then Z_{Nk} and $1-Z_{Nk}$ tend toward definite limits which we shall denote by Z_k and $1-Z_k$ respectively. Let us add to the subset the $(k+1)$ -th watch. In the game based on this new subset the "change" will occur in two subsequent lots in two cases:

(1) if in the game based on the subset of the first k watches there was "change" and on the $(k+1)$ -th watch "no change"; because then in the first two subsequent lots we have in the subset of the first k watches an odd number of "blacks" and in the second lot an even number of "blacks" or *vice versa*; and on the $(k+1)$ -th we have two "blacks" or two "whites;"

(2) if in the game based on the subset of the first k watches there was "no change" and on the $(k+1)$ -th watch there was "change"; because then in both subsequent lots in the subset of the first k watches we have an even or odd number of "blacks," and on the $(k+1)$ -th watch we have the combinations, "black-white or "white-black."

Taking into consideration "the independence" of the occurrence of "change" and "no change" on particular watches we find for the limit of frequency of "change"

$$Z_{k+1} = Z_k(1-\delta_{k+1}) + (1-Z_k)\delta_{k+1}.$$

Analogously we obtain $1-Z_{k+1} = (1-Z_k)(1-\delta_{k+1}) + Z_k\delta_{k+1}$.

It follows that $2Z_{k+1}-1 = (1-2\delta_{k+1})(2Z_k-1)$

and by means of full induction

$$1-2Z_m = (1-2\delta_1)(1-2\delta_2)\dots(1-2\delta_m).$$

As $|1-2\delta_k| \leq \lambda < 1$, where λ is a constant number, we obtain

$$\lim_{m \rightarrow \infty} (1-2Z_m) = 0, \text{ or } \lim_{m \rightarrow \infty} Z_m = \frac{1}{2}.$$

MECHANISTIC MODEL OF A RANDOM PHENOMENON

We can prove now, that the frequency of combinations AA , BB , AB and BA in two subsequent lots of the game, based on all m watches approaches the limit $\frac{1}{4}$, when $N \rightarrow \infty$ and $m \rightarrow \infty$. First of all we have $(AB)_{Nm} + (BA)_{Nm} = Z_{Nm}$. From this follows that

$$\lim_{N \rightarrow \infty} ((AB)_{Nm} + (BA)_{Nm}) = Z_m.$$

Moreover we have

$$\lim_{N \rightarrow \infty} ((AB)_{Nm} + (AA)_{Nm}) = (A)_m \quad \text{and} \quad \lim_{N \rightarrow \infty} ((BA)_{Nm} + (AA)_{Nm}) = (A)_m,$$

or

$$\lim_{N \rightarrow \infty} ((AB)_{Nm} - (BA)_{Nm}) = 0.$$

We thus obtain

$$\lim_{N \rightarrow \infty} (AB)_{Nm} = \lim_{N \rightarrow \infty} (BA)_{Nm} = \frac{1}{2} Z_m.$$

As we have proved above, that $\lim_{m \rightarrow \infty} Z_m = \frac{1}{2}$, we obtain

$$\lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (AB)_{Nm} = \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (BA)_{Nm} = \frac{1}{4}.$$

Taking into consideration, that

$$(AA)_{Nm} + (BB)_{Nm} + (AB)_{Nm} + (BA)_{Nm} = 1,$$

it may easily be shown that also

$$\lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (AA)_{Nm} = \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (BB)_{Nm} = \frac{1}{4},$$

or that two subsequent lots of the game based on m watches are asymptotically "independent" of each other to any degree of approximation if m is sufficiently large.

4. GENERALISATIONS

First, we shall deal with the "independence" of the j -th and $(j+a)$ -th lot. Let a be a definite positive integer. Let us introduce the following notations. The frequency of "change" from black to white and *vice versa* between the j and $j+a$ positions of the hand on the k -th watch will be denoted by $\delta_{Nk}^{(a)}$. Then the frequency of the hands being both in the lot j and $j+a$ in the white or black sector will be equal to $1 - \delta_{Nk}^{(a)}$. Moreover, let us denote by A_j and B_j the phenomena consisting of the occurrence of A and B respectively in the j -th lot.

The frequency of occurrence of combinations AA , BB , AB and BA in j and $j+a$ lots of the game based on the first k watches will be denoted by $(A_j A_{j+a})_{Nk}$, $(B_j B_{j+a})_{Nk}$, $(A_j B_{j+a})_{Nk}$ and $(B_j A_{j+a})_{Nk}$; and in the game based on all the m watches by $(A_j A_{j+a})_{Nm}$, $(B_j B_{j+a})_{Nm}$, $(A_j B_{j+a})_{Nm}$ and $(B_j A_{j+a})_{Nm}$. Finally, we shall denote $a\theta_k - [a\theta_k]$ by $\theta_k^{(a)}$.

Theorem 3: If $|1-2p_k| \leq \alpha < 1$ and $\theta_k^{(a)} \geq \beta^{(a)} > 0$, $1-\theta_k^{(a)} \geq \beta^{(a)} > 0$, $|\frac{1}{2}-\theta_k^{(a)}| \geq \beta^{(a)} > 0$ where α and $\beta^{(a)}$ are constant numbers then

$$\begin{aligned} \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (A_j A_{j+a})_{Nm} &= \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (B_j B_{j+a})_{Nm} \\ &= \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (A_j B_{j+a})_{Nm} = \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (B_j A_{j+a})_{Nm} = \frac{1}{4}. \end{aligned}$$

Proof: From Theorem VI of the Appendix the following generalisations of the characteristics (5) and (6) of the positions of the hands may be derived:

(5) $\delta_{Nk}^{(a)}$ approaches a certain limit $\delta_k^{(a)}$ if $N \rightarrow \infty$, or

$$\lim_{N \rightarrow \infty} \delta_{Nk}^{(a)} = \delta_k^{(a)}.$$

Moreover, if $|1-2p_k| \leq \alpha < 1$ and $\theta_k^{(a)} \geq \beta^{(a)} > 0$, $1-\theta_k^{(a)} \geq \beta^{(a)} > 0$, $|\frac{1}{2}-\theta_k^{(a)}| \geq \beta^{(a)} > 0$ where α and $\beta^{(a)}$ are constant numbers, then $|1-2\delta_k^{(a)}| \leq \lambda^{(a)} < 1$ where $\lambda^{(a)}$ is as well a constant number.

(6) The frequency of the occurrence of "change" (or "no change") in the j -th or $(j+a)$ -th lot on the k -th watch is "independent" of the frequency of "change" (or "no change") on other watches. For instance, the limit of frequency of the combination of "change" on the k -th watch and of "no change" on the l -th watch in the j -th and $(j+a)$ -th lots amounts to $\delta_k^{(a)} (1-\delta_l^{(a)})$.

On the basis of these characteristics the theorem may be demonstrated in the same way as the previous one. A further generalisation concerns the simultaneous independence of $j+1, j+2, \dots, j+a$ lots from the j -th lot.

Let us denote the least multiple of the integers not larger than a by v and $v\theta_k - [v\theta_k]$ by $\theta_k^{(v)}$.

Theorem 4: If $|1-2p_k| \leq \alpha < 1$ and $\theta_k^{(v)} \geq \beta^{(v)} > 0$, $1-\theta_k^{(v)} \geq \beta^{(v)} > 0$, $|\frac{1}{2}-\theta_k^{(v)}| \geq \beta^{(v)} > 0$ where α and $\beta^{(v)}$ are constant numbers, then

$$\begin{aligned} \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (A_j A_{j+i})_{Nm} &= \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (B_j B_{j+i})_{Nm} = \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (A_j B_{j+i})_{Nm} \\ &= \lim_{m \rightarrow \infty} \lim_{N \rightarrow \infty} (B_j A_{j+i})_{Nm} = \frac{1}{4} \quad (i = 1, 2, \dots, a). \end{aligned}$$

Proof: It follows from Theorem 3 that Theorem 4 holds good, when conditions (a), (b) and (c) are simultaneously fulfilled

$$(a) \quad j\theta_k - [j\theta_k] \geq \beta^{(i)} > 0,$$

$$(b) \quad 1-j\theta_k + [j\theta_k] \geq \beta^{(i)} > 0,$$

$$(c) \quad |\frac{1}{2}-j\theta_k + [j\theta_k]| \geq \beta^{(i)} > 0,$$

where $\beta^{(i)}$ ($i = 1, 2, \dots, a$) are constant numbers. We shall prove now, that these conditions are fulfilled for $\beta^{(i)} = i \beta^{(v)}/v$.

First of all we have

$$(a') \quad v\theta_k - [v\theta_k] \geq \beta^{(v)} > 0,$$

$$(b') \quad 1-v\theta_k + [v\theta_k] \geq \beta^{(v)} > 0,$$

$$(c') \quad |\frac{1}{2}-v\theta_k + [v\theta_k]| \geq \beta^{(v)} > 0.$$

MECHANISTIC MODEL OF A RANDOM PHENOMENON

Let us start from condition (a). We have obviously

$$i\theta_k - [i\theta_k] > 0$$

(because θ_k is an irrational number). Let us multiply both sides of this inequality by the expression v/i , which is an integer, as v is a multiple of i :

$$v\theta_k - \frac{v}{i} [i\theta_k] > 0.$$

As $[v\theta_k]$ is the largest integer contained in $v\theta_k$ we have :

$$v\theta_k - \frac{v}{i} [i\theta_k] \geq v\theta_k - [v\theta_k].$$

Hence and from (a') it follows that

$$v\theta_k - \frac{v}{i} [i\theta_k] \geq \beta^{(v)}.$$

Dividing by v/i we obtain $i\theta_k - [i\theta_k] \geq \frac{i}{v} \beta^{(v)} > 0$,

which proves that condition (a) is fulfilled.

Let us consider now condition (b). We have $1 - i\theta_k + [i\theta_k] > 0$. Let us multiply both sides of this inequality by v/i

$$\frac{v}{i} - v\theta_k + \frac{v}{i} [i\theta_k] > 0.$$

Since $[v\theta_k] + 1$ is the smallest integer greater than $v\theta_k$, we have

$$\frac{v}{i} - v\theta_k + \frac{v}{i} [i\theta_k] \geq 1 - v\theta_k + [v\theta_k].$$

Hence and from (b') it follows that

$$\frac{v}{i} - v\theta_k + \frac{v}{i} [i\theta_k] \geq \beta^{(v)} > 0;$$

Dividing by v/i we obtain $1 - i\theta_k + [i\theta_k] \geq \frac{i}{v} \beta^{(v)} > 0$,

which proves that condition (b) is fulfilled.

Finally, for condition (c) the proof is as follows. Multiplying the expression $|\frac{1}{2} - i\theta_k + [i\theta_k]|$ by v/i we obtain :

$$\left| \frac{v}{2i} - v\theta_k + \frac{v}{i} [i\theta_k] \right|.$$

Since either $\frac{1}{2} + [v\theta_k]$, or $1 + [v\theta_k]$ is the multiple of the number $\frac{1}{2}$ closest to $v\theta_k$, we have either

$$\left| \frac{v}{2i} - v\theta_k + \frac{v}{i} [i\theta_k] \right| \geq \left| \frac{1}{2} - v\theta_k + [v\theta_k] \right|,$$

or

$$\left| \frac{v}{2i} - v\theta_k + \frac{v}{i} [i\theta_k] \right| \geq 1 - v\theta_k + [v\theta_k].$$

But according to the assumptions (b') and (c') the expressions on the right hand of both these inequalities are not smaller than $\beta^{(v)}$, which is positive. Thus

$$\left| \frac{v}{2i} - v\theta_k + \frac{v}{i} [i\theta_k] \right| \geq \beta^{(v)} > 0$$

and after dividing by v/i we obtain

$$|\frac{1}{2} - i\theta_k + [i\theta_k]| \geq \frac{i}{v} \beta^{(v)} > 0,$$

which proves that condition (c) is fulfilled as well.

Theorem 4 may be formulated also as follows:

The result of the watch game is to any degree of approximation "independent" of the results of the preceding a lots when N and m are sufficiently large, if $\theta_k^{(v)}$, $1 - \theta_k^{(v)}$ and $|\frac{1}{2} - \theta_k^{(v)}|$ are greater than a certain positive constant, where $\theta_k^{(v)} = v\theta_k - [v\theta_k]$ and v is the least multiple of integers, not greater than a .

It should be noticed that the limiting conditions for the set of θ_k depend on a . The same concerns the magnitude of m necessary for the achievement of a given degree of approximation to "independence" of the results of the preceding a lots.

As will be seen from the above we may make the results of the watch game as close as we wish to a random sequence. First of all we assume any large, but definite a . Let us next choose a set of m of arithmetically independent irrational numbers $\varphi_1, \varphi_2, \dots, \varphi_m$ satisfying the conditions :

$$\varphi_k - [\varphi_k] > \beta^{(v)}, \quad 1 - \varphi_k + [\varphi_k] > \beta^{(v)} \quad \text{and} \quad |\frac{1}{2} - \varphi_k + [\varphi_k]| > \beta^{(v)},$$

where $\beta^{(v)}$ is a constant positive number. Let us assume that the velocities of the points of the hands of m watches are the numbers $\varphi_1/v, \varphi_2/v, \dots, \varphi_m/v$ where v is the least multiple of integers which are smaller than or equal to a . If m is sufficiently large the result of the lot will be "independent" of the results of the preceding a lots to any degree of approximation.

MECHANISTIC MODEL OF A RANDOM PHENOMENON

It is still worth to consider a particular case of the set of θ_k fulfilling the conditions

$$(a') \quad v\theta_k - [v\theta_k] > \beta^{(v)} > 0,$$

$$(b') \quad 1 - v\theta_k + [v\theta_k] > \beta^{(v)} > 0,$$

$$(c') \quad |\frac{1}{2} - v\theta_k + [v\theta_k]| > \beta^{(v)} > 0.$$

It may be proved that these conditions are satisfied when $\theta_1, \theta_2, \dots, \theta_m$ are contained in a sufficiently small interval. Let us denote by $\epsilon^{(v)}$ the length of this interval. We have then $|\theta_k - \theta_1| \leq \epsilon^{(v)}$. Let $\epsilon^{(v)}$ be sufficiently small to fulfill conditions:

$$(e) \quad v\epsilon^{(v)} < v\theta_1 - [v\theta_1],$$

$$(f) \quad v\epsilon^{(v)} < 1 - v\theta_1 + [v\theta_1],$$

$$(g) \quad v\epsilon^{(v)} < |\frac{1}{2} - v\theta_1 + [v\theta_1]|.$$

Hence it follows, that

$$|v\theta_k - v\theta_1| < v\theta_1 - [v\theta_1] \text{ and } |v\theta_k - v\theta_1| < 1 - v\theta_1 + [v\theta_1],$$

and in consequence

$$[v\theta_k] = [v\theta_1].$$

We obtain further

$$v\theta_k - [v\theta_k] = v\theta_1 - [v\theta_1] + v(\theta_k - \theta_1),$$

$$1 - v\theta_k + [v\theta_k] = 1 - v\theta_1 + [v\theta_1] - v(\theta_k - \theta_1),$$

$$|\frac{1}{2} - v\theta_k + [v\theta_k]| = |\frac{1}{2} - v\theta_1 + [v\theta_1] - v(\theta_k - \theta_1)|.$$

It follows from the condition $|\theta_k - \theta_1| \leq \epsilon^{(v)}$ that

$$v\theta_k - [v\theta_k] \geq v\theta_1 - [v\theta_1] - v\epsilon^{(v)},$$

$$1 - v\theta_k + [v\theta_k] \geq 1 - v\theta_1 + [v\theta_1] - v\epsilon^{(v)},$$

$$|\frac{1}{2} - v\theta_k + [v\theta_k]| \geq |\frac{1}{2} - v\theta_1 + [v\theta_1]| - v\epsilon^{(v)}.$$

From conditions (e), (f) and (g) for $\epsilon^{(v)}$ it follows, that the expressions on the right hand side of the inequality are positive numbers. If we denote the smallest of them by $\beta^{(v)}$ we obtain the characteristics (a'), (b') and (c').

As will be seen from the above, the crowding of velocities θ_k within a sufficiently small interval dependent on a and a sufficiently large number of m watches create conditions, under which the result of a lot is independent of the results of preceding a lots to any degree of approximation. It will be easily seen, that such conditions could also be assured by crowding of θ_k into a finite number of sufficiently small intervals.

It may be of interest to note that the results arrived at in this paper have been applied in the work of the Institute of Mathematical Machines, Warsaw. By means of a programme for an electronic computer corresponding approximately to the mechanism considered here, a quasi-random sequence will be generated which is required for the application of the so-called Montecarlo method.

Appendix

Certain characteristics of the sequences of irrational numbers $\theta, 2\theta, 3\theta, \dots$

Definition 1 : Let a sequence of real numbers be given

$$f(1), f(2), \dots, f(N).$$

Let us denote by $N(\gamma)$ the number of members fulfilling the conditions

$$f(k) - [f(k)] < \gamma \text{ where } 0 \leq \gamma \leq 1.$$

We shall say, that the sequence f has equipartition, if

$$\lim_{N \rightarrow \infty} \frac{N(\gamma)}{N} = \gamma$$

for all admissible values of γ .

Definition 2: Let there be m sequences

$$f_1(1), f_1(2), \dots, f_1(N),$$

$$f_2(1), f_2(2), \dots, f_2(N),$$

$$\dots \dots \dots$$

$$f_m(1), f_m(2), \dots, f_m(N).$$

Let us denote by $N(\gamma_1, \gamma_2, \dots, \gamma_m)$ the number of columns of this set of sequences, which for each k from 1 to m satisfy the condition

$$f_k(j) - [f_k(j)] < \gamma_k,$$

where $0 \leq \gamma_k < 1$. We shall say, that the set of sequences $f_1, f_2, \dots, f_m$ has equipartition, if

$$\lim_{N \rightarrow \infty} \frac{N(\gamma_1, \gamma_2, \dots, \gamma_m)}{N} = \gamma_1 \gamma_2 \dots \gamma_m$$

for all admissible values of the γ 's. If the set of sequences $f_1, f_2, \dots, f_m$ has equipartition, then each of its subsets has it as well. We can show it directly by imposing condition $\gamma = 1$ on all f outside the subset.

Theorem I : If θ is any given irrational number, then the sequence

$$\theta, 2\theta, 3\theta, \dots, N\theta \quad \dots \quad (1)$$

has equipartition.

This theorem was proved in 1909 by Sierpinski and by Bohl. In 1916 Weyl offered a simple proof based on a method having a wider application.*

* See J. F. Koksma. *Diophantische Approximationen*, Berlin, 1936, p. 92.

MECHANISTIC MODEL OF A RANDOM PHENOMENON

Theorem II : (A particular case of Weyl's other theorem).^{*} *The set of sequences*

$$\begin{aligned} &\theta_1, 2\theta_1, \dots, N\theta_1, \\ &\theta_2, 2\theta_2, \dots, N\theta_2, \qquad \dots \quad (2) \\ &\dots \dots \dots \\ &\theta_m, 2\theta_m, \dots, N\theta_m \end{aligned}$$

has equipartition, if for any given set of integers $h_1, h_2, \dots, h_m$ which are not simultaneously equal to zero, the expression

$$h_1\theta_1 + h_2\theta_2 + \dots + h_m\theta_m$$

is an irrational number.

Equivalent of Theorem II : If all the numbers $\theta_1, \theta_2, \dots, \theta_m$ are irrational and arithmetically independent, then the set (2) (and each of its subsets) has equipartition.

From Theorem I we derive the following :

Theorem III : *Let us denote by S the number of terms of the sequence (1), which satisfy the condition (a) or (b) :*

$$(a) \quad j\theta - [j\theta] < \gamma, \quad (j+1)\theta - [(j+1)\theta] \geq \gamma,$$

$$(b) \quad j\theta - [j\theta] \geq \gamma, \quad (j+1)\theta - [(j+1)\theta] < \gamma.$$

The ratio S/N approaches the limit determined subsequently by formulae (3), (4) and (5) when $N \rightarrow \infty$.

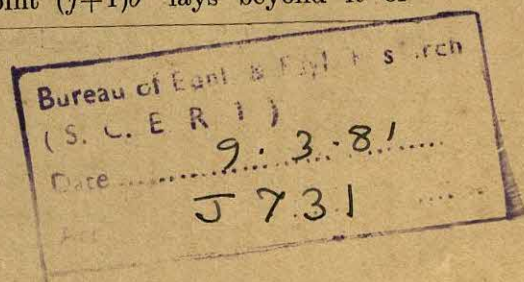
Let us denote $\theta - [\theta]$ by θ' . It will be easily seen, that

$$j\theta - [j\theta] = j\theta' - [j\theta'].$$

Therefore in formulae (a) and (b) it is possible to substitute θ' , which is less than one, for θ . It may be also assumed that $\gamma \leq \frac{1}{2}$ without impinging on the generality of the proof.

The theorem may be easily demonstrated graphically. Let us measure on the circle of a circumference 1 the arc of the length $j\theta'$ from the zero point clockwise. $j\theta' - [j\theta']$ is then the distance of point $j\theta'$ from zero. The conditions (a) and (b) mean, that should point $j\theta'$ lay within the arc γ , the point $(j+1)\theta'$ lays beyond it or

^{*} Ibidem, p. 93.



vice versa. Without impinging on the generality of the proof it may be assumed, that the beginning of the arc γ is the zero point. Let us consider three cases

- (A) $\theta' < \gamma$,
 (B) $\gamma \leq \theta' < 1-\gamma$,
 (C) $\theta' \geq 1-\gamma$,

to which correspond the figures 1, 2 and 3. As we shall prove below the points $j\theta'$ laying on the arcs marked by a heavy line satisfy the conditions (a) and (b).

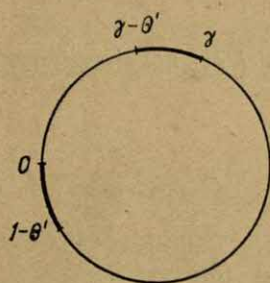


Fig. 1

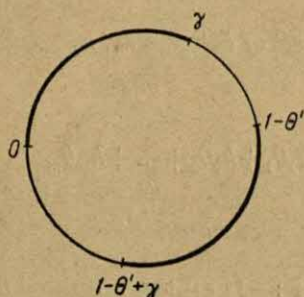


Fig. 2

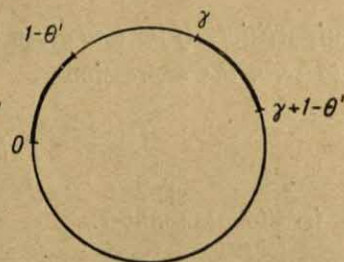


Fig. 3

Let us consider the case (A). If the point $j\theta'$ lays within the arc $0, \gamma$, the point $(j+1)\theta'$ lays beyond it then and then only when the point $j\theta'$ lays on the arc $\gamma-\theta', \gamma$. If the point $j\theta'$ lays within the arc $\gamma, 0$, the point $(j+1)\theta'$ lays beyond it then and then only when $j\theta'$ lays on the arc $1-\theta', 0$.

It follows from Theorem I, that in this case

$$\lim_{N \rightarrow \infty} \frac{S}{N} = 2\theta'. \quad \dots (3)$$

In the case (B), if the point $j\theta'$ lays within the arc $0, \gamma$, then the point $(j+1)\theta'$ lays beyond it. If $j\theta'$ lays within the arc $\gamma, 0$ the point $(j+1)\theta'$ lays beyond then and then only when the point $j\theta'$ lays on the arc $1-\theta', 1-\theta'+\gamma$. In this case we have

$$\lim_{N \rightarrow \infty} \frac{S}{N} = 2\gamma. \quad \dots (4)$$

Finally, in the case (C), if the point $j\theta'$ lays within the arc $0, \gamma$, then the point $(j+1)\theta'$ lays beyond it then and then only when the point $j\theta'$ lays on the arc $0, 1-\theta'$. If $j\theta'$ lays within the arc $\gamma, 0$, the point $(j+1)\theta'$ lays then and then only beyond it, when $j\theta'$ lays on the arc $\gamma, \gamma+1-\theta'$. In this case we obtain

$$\lim_{N \rightarrow \infty} \frac{S}{N} = 2(1-\theta'). \quad \dots (5)$$

Thus we have proved, that when $N \rightarrow \infty$ then the relation of the number S of terms of the sequence $\theta, 2\theta, \dots, N\theta$, satisfying the conditions (a) or (b) to the total number

MECHANISTIC MODEL OF A RANDOM PHENOMENON

of terms approaches the limit defined by the formulae (3), (4) and (5), which we denote by δ :

$$\lim_{N \rightarrow \infty} \frac{S}{N} = \delta.$$

From the Theorems II and III we deduce the following

Theorem IV : *Let us denote by $S_{1,2,\dots,N}$ the number of columns of the set of sequences (2) satisfying for each k from 1 to m the conditions (a) or (b) :*

$$(a) \quad j\theta_k - [j\theta_k] < \gamma, \quad (j+1)\theta_k - [(j+1)\theta_k] \geq \gamma,$$

$$(b) \quad j\theta_k - [j\theta_k] \geq \gamma, \quad (j+1)\theta_k - [(j+1)\theta_k] < \gamma.$$

Moreover let us denote the number of terms satisfying conditions (a) or (b) in the sequence k by S_k and $\lim (S_k/N)$ by δ_k . We have then

$$\lim_{N \rightarrow \infty} \frac{S_{1,2,\dots,N}}{N} = \delta_1 \delta_2 \dots, \delta_m.$$

This will hold good also for each subset of the set of sequences (2). Let us imagine m circles with circumferences equal to 1. On the k -th circle points $j\theta_k$ satisfying condition (a) or (b) lay on two definite arcs of joint length δ_k . From Theorem II it follows, that the ratio of the number of these j , for which conditions (a) or (b) are satisfied on all circles, i.e. $S_{1,2,\dots,N}$, to N is equal to the product $\delta_1 \delta_2 \dots, \delta_m$.

Theorem V : *If $|1-2\gamma_k| \leq \alpha < 1$ and $\theta'_k \geq \beta > 0$, $1-\theta'_k \geq \beta > 0$, $|\frac{1}{2}-\theta'_k| \geq \beta > 0$ where α and β are constant numbers, then $|1-2\delta_k| \leq \lambda < 1$, where λ is also a constant number.*

Proof : First of all it should be pointed out, that without impinging on the generality of the theorem we may assume $\alpha > \frac{1}{2}$ and $\beta < \frac{1}{4}$. For the three cases considered in the proof of the Theorem III we have for δ_k and $|1-2\delta_k|$ (assuming like in the other proof, that $\gamma_k \leq \frac{1}{2}$)

case	δ_k	$ 1-2\delta_k $
(A) $\theta'_k < \gamma_k \leq \frac{1}{2}$	$2\theta'_k$	$ 1-4\theta'_k $
(B) $\gamma_k \leq \theta'_k < 1-\gamma$	$2\gamma_k$	$ 1-4\gamma_k $
(C) $\frac{1}{2} \leq 1-\gamma_k \leq \theta'_k$	$2(1-\theta'_k)$	$ 1-4(1-\theta'_k) $

In the cases (A) and (C) from conditions

$$\theta'_k \geq \beta < \frac{1}{4}, \quad 1-\theta'_k \geq \beta < \frac{1}{4}, \quad |\frac{1}{2}-\theta'_k| \geq \beta < \frac{1}{4}$$

it follows, as will be easily seen, that

$$|1-2\delta_k| \leq 1-4\beta < 1.$$

In the case (B) it follows from condition $|1-2\gamma_k| \leq \alpha$ and from the inequality $\gamma_k \leq \frac{1}{2}$ that

$$\gamma_k \geq \frac{1}{2}(1-\alpha).$$

Next it follows from condition $|\frac{1}{2}-\theta'| \geq \beta$, that

$$\text{either } \theta'_k \leq \frac{1}{2}-\beta \text{ or } 1-\theta'_k \leq \frac{1}{2}-\beta.$$

In addition we have $\gamma_k \leq \theta'_k$ and $1-\gamma_k > \theta'_k$ i.e. $\gamma_k < 1-\theta'_k$.

Hence
$$\gamma_k \leq \frac{1}{2}-\beta.$$

Thus we have
$$\frac{1}{2}(1-\alpha) \leq \gamma_k \leq \frac{1}{2}-\beta.$$

Hence we obtain for the case $\gamma_k \leq \frac{1}{4}$

$$|1-4\gamma_k| \leq 1-4(1-\alpha)/2 = 2\alpha-1 < 1.$$

and for the case $\gamma_k > \frac{1}{4}$

$$|1-4\gamma_k| \leq 1-4\beta < 1.$$

From this follows that if λ is equal to the greater of the two numbers $2\alpha-1$ and $1-4\beta$, then we have

$$|1-2\delta_k| \leq \lambda < 1.$$

Theorem VI : *Theorems III, IV and V hold good, if conditions (a) and (b) take on a more general form*

$$(a) \quad j\theta_k - [j\theta_k] < \gamma_k, \quad (j+a)\theta_k - [(j+a)\theta_k] \geq \gamma_k,$$

$$(b) \quad j\theta_k - [j\theta_k] \geq \gamma_k, \quad (j+a)\theta_k - [(j+a)\theta_k] < \gamma_k,$$

where a is a positive integer and if in Theorem V θ' will be replaced by $\theta_k^{(a)} = a\theta_k - [a\theta_k]$.

These theorems may be proved as above by replacing θ'_k by $\theta_k^{(a)}$.

Paper received : December, 1959.

A STUDY OF LARGE SAMPLE TEST CRITERIA THROUGH PROPERTIES OF EFFICIENT ESTIMATES

PART I: TESTS FOR GOODNESS OF FIT AND CONTINGENCY TABLES

By C. RADHAKRISHNA RAO

Indian Statistical Institute

SUMMARY. The existence of m.l. and m.l. equation estimates and their efficiency, in the case of sampling from a finite multinomial distribution, have been established under conditions weaker than those assumed by earlier writers. For instance, the existence of the second derivatives of the cell probabilities as functions of the parameters is not necessary. The asymptotic distributions of the chi-square goodness of fit test and test criteria for examining composite hypotheses in contingency tables have been shown to depend mainly on the efficiency (in the new sense) of estimates used in the construction of the test criteria. The m.l. estimate satisfies the efficiency condition and is, in the sense of second order efficiency, more suitable for the construction of test criteria than other efficient estimates. A new test has been proposed for examining the expected frequencies computed under two different hypotheses of a special nature.

1. INTRODUCTION

In two previous papers (Rao, 1960a, 1960b), the author gave a new formulation of the concept of asymptotic efficiency of estimates of parameters and established an optimum property (second order efficiency) of the maximum likelihood (m.l.) estimate. This shows that although the class of estimation procedures providing efficient estimates is very wide, the m.l. method is distinguishable from the rest by its maximum second order efficiency. It has been observed that m.l. estimates (which are efficient in the new sense) provide a good summary of data in large samples in the sense that inference based on these estimates is equivalent to that obtained by utilizing the whole data. The approach to the problem of estimation and the specific results obtained have been based on the concepts introduced by Fisher (1921, 1925).

It is also well known that in the construction of large sample test criteria efficient estimates play an important role. For instance, in obtaining a χ^2 goodness of fit, the expected values are calculated by inserting efficient estimates such as those obtained by maximising likelihood or minimising χ^2 . The object of the present series of articles is to show how the concept of efficiency in the new sense plays a key role in the construction of large sample criteria and in the derivation of their asymptotic distributions. It is also suggested that preference should be given to m.l. estimates, because of its maximum second order efficiency, among all efficient estimates in the construction of large sample criteria.

Incidentally, the existence of the m.l. estimates, their consistency, efficiency and asymptotic normality of distribution have all been deduced, in the case of the finite multinomial under conditions weaker than those assumed by Cramér (1946);

for instance, without assuming second order differentiability of the cell probabilities as functions of unknown parameters. The χ^2 goodness of fit criterion is shown to have the usual asymptotic distribution even under these weaker conditions.

2. DEFINITIONS AND ASSUMPTIONS

Let $\pi_1, \dots, \pi_k$, the probabilities in the k cells of a multinomial distribution, depend on a vector $\theta = (\theta_1, \dots, \theta_q)$ of parameters belonging to a q dimensional real space R^q or to a nondegenerate interval of R^q . If $p_1, \dots, p_k$ are the observed proportions in the k cells for a sample of size n , the log likelihood of θ is proportional to

$$p_1 \log \pi_1 + \dots + p_k \log \pi_k.$$

When π_i are differentiable functions of θ , the maximum likelihood (m.l.) equations are defined by

$$z_r = \sum \frac{p_i}{\pi_i} \frac{\partial \pi_i}{\partial \theta_r} = 0, \quad r = 1, \dots, q. \quad \dots (2.1)$$

The information matrix, which is the variance covariance matrix of z_r , is denoted by $\Lambda = BB' = (i_{rs})$, where the matrix B and i_{rs} are defined as follows:

$$B = \left(\frac{1}{\sqrt{\pi_i}} \frac{\partial \pi_i}{\partial \theta_r} \right), \quad i_{rs} = \sum \frac{1}{\pi_i} \frac{\partial \pi_i}{\partial \theta_r} \frac{\partial \pi_i}{\partial \theta_s}. \quad \dots (2.2)$$

Unless otherwise stated, the variables z_i and the elements i_{rs} denote the values at $\theta_0 = (\theta_1^0, \dots, \theta_q^0)$, the true value of the parameter supposed to be an interior point of the admissible region of θ , such that $\pi_i(\theta_0) > 0$ for each i .

We introduce the following assumptions.

Assumption 1: Every π_i has continuous first derivatives $\partial \pi_i / \partial \theta_s$ at the true value θ_0 .

Assumption 2a: Given a $\delta > 0$, it is possible to find an $\epsilon > 0$, such that

$$\inf_{(\theta - \theta_0)^2 \geq \delta} \sum \pi_i(\theta_0) \log \frac{\pi_i(\theta)}{\pi_i(\theta_0)} \geq \epsilon. \quad \dots (2.3)$$

Assumption 2b: $\pi_i(\theta) \neq \pi_i(\beta)$ for at least one i , when $\theta \neq \beta$, which is the identifiability condition.

It is easy to see that Assumption 2b $\implies$ Assumption 2a, when the admissible interval of θ is closed and $\pi_i(\theta)$ are continuous functions of θ . If the interval is open, it may happen that

$$|\pi_i(\theta_n) - \pi_i(\theta_0)| \rightarrow 0, \text{ for each } i \text{ as } \theta_n \rightarrow \beta \neq \theta_0$$

Assumption 2a prevents such a thing happening and is in a certain sense a strong identifiability condition.

LARGE SAMPLE TEST CRITERIA

Assumption 3: There exists an estimate $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_q)$ such that

- i) $\text{plim } \hat{\theta} = \theta_0$
- ii) $\sqrt{n}d_i = \sqrt{n}(\hat{\theta}_i - \theta_i^0)$ has an asymptotic distribution (a.d.).
- iii) The estimate $\hat{\theta}$ is efficient in the new sense (Rao, 1960a, 1960b), i.e.,

$$\text{plim } [\sqrt{n}(z_r - i_{r1}d_1 - \dots - i_{rq}d_q)] = 0, \quad r = 1, \dots, q \quad \dots (2.4)$$

or in matrix notation

$$\text{plim } (Z - \Lambda D) = 0$$

where Z and D denote the column vectors of the elements $\sqrt{n}z_r$ and $\sqrt{n}d_s$ respectively. We may, in theory, allow $\hat{\theta}$, itself to depend on the unknown true value θ_0 , but it is necessary for the application of the results derived that $\hat{\pi}_i = \pi_i(\hat{\theta})$, the estimate of $\pi_i(\theta)$ should be independent of the assumed true value.

Assumption 3 may appear as a blanket assumption but is deliberately introduced to demonstrate the key role played by the efficiency condition (2.4). Whatever may be the method by which $\hat{\theta}$ is obtained the propositions considered in the paper are valid provided only it satisfies the Assumption 3.

We will, in fact, show in Section 3.5 that Assumptions 1 and 2b imply Assumption 3 when the rank of Λ is q , and Assumptions 1 and 2a imply the stronger result that m.l. estimates exist and satisfy the conditions of Assumption 3 when the rank of Λ is q . The conditions under which the results relating to existence of m.l. estimates, their consistency, efficiency (implying asymptotic normality of distribution) are obtained are much weaker than those considered by earlier writers. It can be proved that under the same or similar less restrictive conditions other methods of estimation such as minimum χ^2 (Neyman, 1949; Rao, 1955a), modified minimum χ^2 (Neyman, 1949), minimum discrepancy (Haldane, 1951), minimum Hellinger distance (Rao, 1960b), also provide estimates satisfying the Assumption 3.

Situations, however, exist where the Assumption 3 is satisfied without even the rank of Λ being q , so that the treatment of the paper is, in some respects, more general.

We have not made any assumption about the rank of Λ at the true value θ_0 and the proofs adopted are valid whatever may be the rank of Λ . If the rank is full, i.e., equal to q , then the Assumption 3 is equivalent to

$$\text{plim}_{n \rightarrow \infty} [\sqrt{n}(d_r - i^{r1}z_1 - \dots - i^{rq}z_q)] = 0 \quad \dots (2.5)$$

and the proofs of the propositions considered are extremely simple when (2.5) holds.

It has been shown in (Rao, 1960b) that the rate of convergence in (2.4) is highest when $\hat{\theta}$ is an m.l. estimate, which indicates some merit in using an m.l. estimate in preference to any other efficient estimate.

3. CHI-SQUARE GOODNESS OF FIT AND ASSOCIATED PROBLEMS

3.1. *Preliminary lemmas.* We shall assemble, to begin with, the key notations and their relationships to be used in the lemmas. The transpose of a column vector is indicated by a prime. The probabilities π_i and their derivatives are taken at the true value θ_0 of the parameter unless stated otherwise.

$$V' = \left(\sqrt{n} \frac{p_1 - \pi_1}{\sqrt{\pi_1}}, \dots, \sqrt{n} \frac{p_k - \pi_k}{\sqrt{\pi_k}} \right)$$

$$U' = \left(\sqrt{n} \frac{\dot{\pi}_1 - \pi_1}{\sqrt{\pi_1}}, \dots, \sqrt{n} \frac{\dot{\pi}_k - \pi_k}{\sqrt{\pi_k}} \right)$$

$$Z' = (\sqrt{n} z_1, \dots, \sqrt{n} z_q), \quad z_r = \sum \frac{p_i}{\pi_i} \frac{\partial \pi_i}{\partial \theta_r}$$

$$D' = (\sqrt{n} d_1, \dots, \sqrt{n} d_q), \quad d_i = \dot{\theta}_i - \theta_i^0$$

$B(q \times k) = \left(\frac{1}{\sqrt{\pi_i}} \frac{\partial \pi_i}{\partial \theta_r} \right)$ so that $BB' = \Lambda$, the dispersion matrix of Z and $Z = BV$.

Observe that $U \sim B'D$, since

$$\sqrt{n} \frac{\dot{\pi}_i - \pi_i}{\sqrt{\pi_i}} \sim \sum \frac{\sqrt{n}}{\sqrt{\pi_i}} \frac{\partial \pi_i}{\partial \theta_r} d_r$$

which follows from the continuity of the derivatives of $\pi_i(\theta)$, where $\sim$ stands for equivalence of asymptotic distributions.

By $\text{cov}(X, Y)$, where X and Y are vectors, is meant the matrix of covariances

$$\begin{pmatrix} \text{cov}(x_1, y_1) & \dots & \text{cov}(x_1, y_k) \\ \dots & \dots & \dots \\ \text{cov}(x_m, y_1) & \dots & \text{cov}(x_m, y_k) \end{pmatrix}$$

For instance, by straight computation, we find

$$\text{cov}(V, Z) = B'.$$

Lemma 1: *The a.d. (asymptotic distribution) of Z is q -variate normal with dispersion matrix Λ .*

This result is a consequence of every linear function $Z'L$, (L being a non-random vector) having an asymptotic normal distribution (Wald and Wolfowitz, 1944). The distribution is, however, singular when the rank of Λ is not full.

LARGE SAMPLE TEST CRITERIA

Lemma 2a: Under the Assumptions 1 and 3, the a.d. of $D' \Lambda D$ is $\chi^2(q)$ when the rank of Λ is full.

If the rank of Λ is q , then by (2.5)

$$D \sim \Lambda^{-1} Z, \quad D' \Lambda D \sim Z' \Lambda^{-1} Z.$$

The a.d. of $Z' \Lambda^{-1} Z$ is $\chi^2(q)$ in virtue of the result of Lemma 1 and the same is then, true of $D' \Lambda D$.

Lemma 2b: Under the Assumptions 1 and 3, the a.d. of $D' \Lambda D$ is $\chi^2(t)$, where t is the rank of Λ .

There exists a matrix C , non-singular such that $C \Lambda C' = \Delta$, where Δ is diagonal and of order q , but with only t elements > 0 and the rest equal to zero. Denoting by ∇ the matrix obtained by replacing the non-zero elements of Δ by their reciprocals we find¹

$$\begin{aligned} \Lambda &= \Lambda C' \nabla C \Lambda \\ D' \Lambda D &= D' \Lambda C' \nabla^{\frac{1}{2}} \nabla^{\frac{1}{2}} C \Lambda D \\ &\sim Z' C' \nabla^{\frac{1}{2}} \nabla^{\frac{1}{2}} C Z. \end{aligned} \tag{3.1.1}$$

The a.d. of $D' \Lambda D$ is same as that of sum of squares of linear functions $Z' C' \nabla^{\frac{1}{2}}$ of Z . The dispersion matrix of $Z' C' \nabla^{\frac{1}{2}}$ is

$$\nabla^{\frac{1}{2}} C \Lambda C' \nabla^{\frac{1}{2}} = \nabla^{\frac{1}{2}} \Delta \nabla^{\frac{1}{2}} = I_t$$

where I_t is a diagonal matrix with t elements equal to unity and the rest to zero.

Hence the a.d. of (3.1.1) is $\chi^2(t)$ and so is that of $D' \Lambda D$.

Lemma 3: Under the Assumptions 1 and 3 the a.d. of $U'U = n \sum (\pi_i - \pi_i)^2 / \pi_i$ is $\chi^2(t)$ where $t = \text{rank } \Lambda$.

Since $U'U \sim D'BB'D = D' \Lambda D$, the result follows from Lemma 2a or 2b.

Lemma 4: The Assumptions 1 and 3 imply $\text{plim } (V-U)'U = 0$ or writing out in full

$$\text{plim } n \sum \frac{(p_r - \pi_r)(\pi_r - \pi_r)}{\pi_r} = 0.$$

This follows from

$$(V-U)'U = V'B'D - D' \Lambda D = (Z' - D' \Lambda)D \rightarrow 0 \text{ in probability by (2.4).}$$

Lemma 5: Under the same conditions as in Lemma 4, a.c. $(V-U, U) = 0$.

Since $BB' = \Lambda$, there exists a matrix G such that $B = \Lambda G$, and therefore, $U \sim B'D = G' \Lambda D \sim G'Z$. Consider

$$\begin{aligned} \text{a.c. } (V - G'Z, G'Z) &= \text{a.c.}(V, G'Z) - \text{a.c.}(G'Z, G'Z) \\ &= \text{c.}(V, Z)G - G' \Lambda G \\ &= B'G - G' \Lambda G = (B' - G' \Lambda)G = 0. \end{aligned}$$

¹ The matrix $C' \nabla C$ is defined to be a pseudo-inverse of Λ . The properties of such inverses and their use in statistics are discussed by the author (Rao, 1955b). The relationship $\Lambda = \Lambda C' \nabla C \Lambda$ is established by post and pre-multiplying with C' and C and using the relation $C \Lambda C' = \Delta$.

Lemma 6* : Let X be a p -variate normal variable with mean zero and dispersion matrix H of rank $f \leq p$. If $Q_1(X)$ and $Q_2(X)$ are two quadratic forms such that $[Q_1(X) + Q_2(X)]$ is $\chi^2(c)$, $Q_1(X)$ is $\chi^2(b)$ and $Q_2(X)$ is non-negative, then $Q_2(X)$ is $\chi^2(c-b)$.

Transform X to W such that $w_1, \dots, w_f$ are each independently $N(0, 1)$ and $w_{f+1}, \dots, w_p$ are all $N(0, 0)$. The transformed quadratic forms $Q_1(W)$, $Q_2(W)$ are, therefore, equivalent, with probability one, to those obtained by omitting $w_{f+1}, \dots, w_p$. The problem reduces to the case of $G = I$ and $p = f$.

Consider an orthogonal transformation from X to Y such that

$$X'X = Y'Y \text{ and } Q_1(X) + Q_2(X) = \sum \lambda_i y_i^2.$$

Since $Q_1(X) + Q_2(X)$ is $\chi^2(c)$, it follows that $\lambda_i = 1$, $i \leq c$ and zero otherwise. Hence

$$y_1^2 + \dots + y_c^2 = Q_1(Y) + Q_2(Y). \quad \dots (3.1.2)$$

Since $Q_2(Y)$ is non-negative and $Q_1(Y)$ is non-negative being distributed as χ^2 , $Q_1(Y)$ and $Q_2(Y)$ contain only the variables $y_1, \dots, y_c$. The problem is thus reduced to the case of c independent normal variables, each with zero mean and unit variance, together with the condition (3.1.2). Under such conditions, given that $Q_1(Y)$ is $\chi^2(b)$, it follows that $Q_2(Y)$ is $\chi^2(c-b)$, as may be proved by considering an orthogonal transformation from Y to Z such that

$$\sum y_i^2 = \sum z_i^2 \text{ and } Q_1(Y) = z_1^2 + \dots + z_b^2.$$

3.2. The goodness of fit test.

Theorem 1: Under the Assumptions 1 and 3, the a.d. of

$$\sum \frac{n(p_r - \pi_r)^2}{\pi_r}$$

is $\chi^2(k-t-1)$, where $t = \text{rank } \Lambda$.

Consider

$$\sum \frac{n(p_r - \pi_r)^2}{\pi_r} = \sum \frac{n(p_r - \pi_r)^2}{\pi_r} + \sum \frac{n(\pi_r - \pi_r)^2}{\pi_r} + \text{product terms}$$

or in matrix notation,

$$V'V \sim (V-U)'(V-U) + U'U \quad \dots (3.2.1)$$

since the product term $\rightarrow 0$ in probability by Lemma 4. The a.d. of $V'V$ is $\chi^2(k-1)$ and that of $U'U$ is $\chi^2(t)$ by Lemma 3. Hence by an application of Lemma 6, $(V-U)'(V-U)$ is $\chi^2(k-1-t)$. Alternatively, since a.c. $(V-U, U) = 0$ by Lemma 5, $(V-U)'(V-U)$ and $U'U$ are asymptotically independently distributed. Hence the a.d. of $(V-U)'(V-U)$ is $\chi^2(k-1-t)$. In order to prove the result of Theorem 1 we observe that

$$\sum \frac{n(p_r - \pi_r)^2}{\pi_r} \sim \sum \frac{n(p_r - \pi_r)^2}{\pi_r}$$

because of continuity of $\pi_r(\theta)$.

* This lemma was proved and included at the suggestion of my colleague, Dr. S. K. Mitra. For proving Theorem 1, we need either Lemma 5 or Lemma 6.

General remarks on the result of Theorem 1. The decomposition of the total χ^2 in (3.2.1) deserves some comments. If the observed frequency in any cell is indicated by $O = np$, the hypothetical frequency by $H = n\pi$ and the estimated (or expected) by $E = n\dot{\pi}$, the equation (3.2.1) may be written

$$\sum \frac{(O-H)^2}{H} \sim \sum \frac{(O-E)^2}{E} + \sum \frac{(E-H)^2}{H} \quad \dots (3.2.2)$$

$$\chi^2(k-1) = \chi^2(k-t-1) + \chi^2(t).$$

For an application of the χ^2 goodness of fit test, it is necessary to enquire whether the rank of Λ is constant over the admissible set of parameters, since the true value of the parameter θ is unknown. What is asserted by Theorem 1 is that the degrees of freedom of a.d. is dependent only on the rank of Λ at the true value, although the rank of Λ may not be same at all points.

The statistic $\Sigma(E-H)^2/H$ may be used to test the hypothesis that a particular set of probabilities is true given that the admissible set has the representation $\pi_i(\theta)$.

3.3. *Test for deviations in any particular set of cells.* Cochran (1954), and Rao and Chakravarty (1956) considered problems where attention is concentrated on the deviation of the observed from the expected frequency in a particular cell or the deviations in a particular set of cells. Tests for examining the singificance of such deviations are extremely useful as they may lead to a suitable explanation of the departure from hypothesis when indicated by a large value of the χ^2 goodness of fit test. The result of the following lemma will be useful in computing the variances and covariances of the deviations.

From Lemma 5 we have

$$\text{a.c. } (V-U, U) = 0 \implies \text{a.c.}(V, U) = \text{a.c.}(U, U).$$

$$\text{Hence a.c. } (V-U, V-U) = \text{c.}(V, V) - \text{a.c.}(U, U)$$

$$\begin{aligned} &= \text{c.}(V, V) - G' \Lambda G \text{ where } G \text{ is as defined in Lemma 4} \\ &= \text{c.}(V, V) - B' \Lambda^{-1} B \text{ when rank } \Lambda = q. \end{aligned} \quad \dots (3.3.1)$$

We have, thus, very simple formulae for finding the asymptotic variances and covariances of the deviations $(p_r - \pi_r)$. For instance, to test the departure in the r -th cell we need

$$\begin{aligned} \text{a.v.}[\sqrt{n}(p_r - \pi_r)] &= \text{v.}[\sqrt{n}(p_r - \pi_r)] - \text{a.v.}[\sqrt{n}(\dot{\pi}_r - \pi_r)] \\ &= \pi_r(1 - \pi_r) - \text{a.v.}[\sqrt{n}(\dot{\pi}_r - \pi_r)]. \end{aligned}$$

When the rank of Λ is q , the last expression can be evaluated in terms of $\Lambda^{-1} = (i^{rs})$ which represents the asymptotic variances and covariances of the estimates $\dot{\theta}_r$

$$\text{a.v.}[\sqrt{n}(\dot{\pi}_r - \pi_r)] = \sum \sum \frac{\partial \pi_r}{\partial \theta_u} \frac{\partial \pi_r}{\partial \theta_v} i^{uv}$$

The deviation $p_r - \pi_r$ can be tested using its standard error.

If the deviations in a number of cells have to be tested simultaneously we have to compute the variance-covariance matrix of the deviations by using the result (3.3.1). With the inverse of the matrix so computed we can construct a quadratic form in the deviations, which has a χ^2 distribution with d.f. equal to the number of deviations to be examined. It may, however, happen that the rank of the variance-covariance matrix of the deviations is not full, in which case the significance of all the deviations cannot be examined simultaneously. We consider only those deviations for which the dispersion matrix remains non-singular.

3.4. *Test for the hypothesis that the parameters belong to a subset.* We need some further assumptions to consider the problem of testing whether the parameter θ belongs to a subset assuming the specification $\pi_i(\theta)$ to be true.

Assumption 4 : The locus of θ in the subset is represented by

$$\theta_i = g_i(\alpha_1, \dots, \alpha_r), \quad i = 1, \dots, q,$$

where g_i admit continuous first derivatives and $(\alpha_1, \dots, \alpha_r)$ is confined to R^r or to some non-degenerate interval in R^r .

Assumption 5: There exist consistent estimates α_i^* , having asymptotic distributions such that

$$\text{plim } \sqrt{n} [y_u - j_{u1}(\alpha_1^* - \alpha_1^0) - \dots - j_{ur}(\alpha_r^* - \alpha_r^0)] = 0, \quad u = 1, \dots, r$$

where $y_u = \frac{\partial}{\partial \alpha_u} \sum p_i \log \pi_i$, and (j_{mp}) is the information matrix for the parameters $\alpha_1, \dots, \alpha_r$.

Assumption 5 is same as Assumption 3 in terms of the new parameters.

Lemma 7 : *The Assumptions 1, 3, 4 and 5 imply*

$$\text{plim } n \sum \frac{(\pi_u^* - \pi_u)(\pi_u^* - \pi_u)}{\pi_u} = 0.$$

By the same argument as in Lemma 4 we have

$$\text{plim } n \sum \frac{(p_u - \pi_u^*)(\pi_u^* - \pi_u)}{\pi_u} = 0$$

and, therefore, Lemma 8 is true if

$$\text{plim } n \sum \frac{(p_u - \pi_u^*)(\pi_u^* - \pi_u)}{\pi_u} = 0. \quad \dots (3.4.1)$$

Since $\sqrt{n}(\pi_u^* - \pi_u) \sim \sqrt{n} \sum \frac{\partial \pi_u}{\partial \alpha_i} (\alpha_i^* - \alpha_i^0)$, to prove (3.4.1) it is enough to show

$$\text{plim } \sum \frac{\sqrt{n}}{\pi_u} \frac{\partial \pi_u}{\partial \alpha_i} (p_u - \pi_u^*) = 0. \quad \dots (3.4.2)$$

LARGE SAMPLE TEST CRITERIA

The left hand side of (3.4.2) is equal to

$$\begin{aligned} & \text{plim} \sum_i \frac{\partial g_i}{\partial \alpha_i} \sum_u \frac{\sqrt{n}}{\pi_u} \frac{\partial \pi_u}{\partial \theta_i} (p_u - \pi_u) \\ &= \sum_i \frac{\partial g_i}{\partial \alpha_i} \text{plim} \sum_u \frac{\sqrt{n}}{\pi_u} \frac{d\pi_u}{d\theta_i} (p_u - \pi_u) \\ &= \sum_i \frac{\partial g_i}{\partial \alpha_i} \text{plim} \sqrt{n}(z_i - i_{i1}d_1 - \dots - i_{iq}d_q) = 0 \text{ by (2.4).} \end{aligned}$$

Theorem 2: Under the Assumptions 1, 3, 4 and 5, the a.d. of

$$n \sum \frac{(\pi_m - \pi_m^*)^2}{\pi_m^*}$$

is $\chi^2(t-a)$, where $t = \text{rank } \Lambda$ and $a = \text{rank}(j_{rs})$.

The result follows from the decomposition

$$n \sum \frac{(\pi_m - \pi_m^*)^2}{\pi_m^*} \sim n \sum \frac{(\pi_m - \pi_m^*)^2}{\pi_m^*} + n \sum \frac{(\pi_m^* - \pi_m)^2}{\pi_m} \dots (3.4.4)$$

where the left hand side is asymptotically $\chi^2(t)$ by Theorem 1, and by Lemma 3, the second term on the right hand side is asymptotically $\chi^2(a)$. Lemma 7 justifies the expansion (3.4.4) and finally an application of Lemma 6 gives the desired result.

General comments on the result of Theorem 2. If we represent by $E_2 = n\pi^*$, expected or estimated frequency under the second hypothesis (that the parameter θ is confined to a subset) the test criterion is

$$\sum \frac{(E_1 - E_2)^2}{E_2} \dots (3.4.5)$$

where it may be noted that E_1 is written for E , the expected frequency under the first specification. Combining the decompositions (3.2.2) and (3.4.4) we have

$$\sum \frac{(O-H)^2}{H} \sim \sum \frac{(O-E_1)^2}{E_1} + \sum \frac{(E_1-E_2)^2}{E_2} + \sum \frac{(E_2-H)^2}{H}. \dots (3.4.6)$$

Some years ago, the author (Rao, 1948) suggested a general criterion from which several large sample tests having asymptotic χ^2 distribution were deduced as special cases. The test (3.4.5) is, however, different from the previous test. Under the null hypothesis both are asymptotically equivalent and so also other tests proposed for the purpose (Mitra, 1955; Rao and Chakravarty, 1956). But differences may exist in their relative efficiencies in finite samples.

3.5. *Sufficient conditions for the validity of the Assumption 3.* The Assumption 3 or 5 as stated may be difficult to verify and it may, therefore, be useful to have some simple conditions under which they are true. The following lemmas provide some answers.

Lemma 8 : *The Assumptions 1 and 2a, together with the non-singularity of Λ imply the Assumption 3.*

We first consider the set S of θ , such that

$$\sum [\pi_i(\theta_0) + \epsilon_1] \log \pi_i(\theta_0) > \sum [\pi_i(\theta_0) - \epsilon_1] \log \pi_i(\theta)$$

where $\epsilon_1 > 0$ is fixed to make $\pi_i(\theta_0) - \epsilon_1 > 0$. For p_i such that $\pi_i(\theta_0) - \epsilon_2 \leq p_i \leq \pi_i(\theta_0) + \epsilon_2$, where $\epsilon_2 < \epsilon_1$,

$$\sum p_i \log \pi_i(\theta_0) \geq \sum [\pi_i(\theta_0) + \epsilon_1] \log \pi_i(\theta_0) > \sum [\pi_i(\theta_0) - \epsilon_1] \log \pi_i(\theta) \geq \sum p_i \log \pi_i(\theta).$$

For θ outside S , and $(\theta - \theta_0)^2 \geq \delta$, $\log \{\pi_i(\theta_0)/\pi_i(\theta)\}$ are bounded in which case

$$\sum p_i \log \frac{\pi_i(\theta_0)}{\pi_i(\theta)}$$

can be made uniformly close (within ϵ difference) to

$$\sum \pi_i(\theta_0) \log \frac{\pi_i(\theta_0)}{\pi_i(\theta)}$$

by choosing p_i sufficiently close to $\pi_i(\theta_0)$, say $\pi_i(\theta_0) - \epsilon_3 \leq p_i \leq \pi_i(\theta_0) + \epsilon_3$. Since by Assumption 2

$$\inf_{(\theta - \theta_0)^2 \geq \delta} \sum \pi_i(\theta_0) \log \frac{\pi_i(\theta_0)}{\pi_i(\theta)} \geq \epsilon > 0$$

$$\sum p_i \log \frac{\pi_i(\theta_0)}{\pi_i(\theta)} > 0 \text{ for all } \theta \text{ such that } (\theta - \theta_0)^2 \geq \delta \quad (3.5.1)$$

when $\pi_i(\theta_0) - \epsilon_4 \leq p_i \leq \pi_i(\theta_0) + \epsilon_4$, where ϵ_4 is smaller of ϵ_2 and ϵ_3 .

Since $\pi_i(\theta_0) > 0$, and $\pi_i(\theta)$ are continuous, the result (3.5.1) shows that the supremum of $\sum p_i \log \pi_i(\theta)$ is attained in the open interval $(\theta - \theta_0)^2 < \delta$. As δ can be chosen arbitrarily small, and p_i is close to $\pi_i(\theta_0)$ with probability one, the value $\hat{\theta}$ at which the supremum is attained provides a consistent estimate of θ . So far we have used only Assumption 2 and continuity of $\pi_i(\theta)$.

If $\pi_i(\theta)$ are differentiable, the derivative of $\sum p_i \log \pi_i(\theta)$ vanishes at $\hat{\theta}$. This shows that the m.l. equation has at least one root which maximises the likelihood and which provides a consistent estimate.

We shall use the condition that rank Λ is q , and the continuity of the derivatives, to prove Assumption 3. The rank of the matrix Λ , however, does not play a significant role in the proof. It is felt that a weaker condition than this may be sufficient for this purpose.

The m.l. equation is

$$\sum \frac{p_r}{\pi_r} \frac{\partial \dot{\pi}_r}{\partial \theta_s} = 0, \quad s = 1, \dots, q$$

$$\text{or } \sum \frac{p - \dot{\pi}_r}{\dot{\pi}_r} \frac{\partial \dot{\pi}_r}{\partial \theta_s} = 0 = \sum \frac{\sqrt{n}(p - \pi_r)}{\pi_r} \frac{\partial \dot{\pi}_r}{\partial \theta_s} - \sum \frac{\sqrt{n}(\dot{\pi}_r - \pi_r)}{\dot{\pi}_r} \frac{\partial \dot{\pi}_r}{\partial \theta_s} \dots \quad (3.5.2)$$

LARGE SAMPLE TEST CRITERIA

The first term on the right hand side of (3.5.2) is $\sim \sqrt{n} z_s$. Using the expansion

$$\sqrt{n}(\dot{\pi}_r - \pi_r) = \Sigma \left(\frac{\partial \pi_r}{\partial \theta_t} + e_{rt} \right) \sqrt{n} d_t, \text{ where } e_{rt} \rightarrow 0 \text{ as } \dot{\theta} \rightarrow \theta_0$$

the second term on the right hand side of (3.5.2) becomes

$$\sqrt{n} \Sigma j_{ts} d_t \quad (3.5.3)$$

where

$$\begin{aligned} j_{ts} &= \Sigma \frac{1}{\pi_r} \frac{\partial \dot{\pi}_r}{\partial \theta_s} \left(\frac{\partial \pi_r}{\partial \theta_t} + e_{rt} \right) \\ &= \Sigma \left(\frac{1}{\pi_r} \frac{\partial \pi_r}{\partial \theta_s} + e'_{rs} \right) \left(\frac{\partial \pi_r}{\partial \theta_t} + e_{rt} \right), e'_{rs} \rightarrow 0 \text{ as } \dot{\theta} \rightarrow \theta_0 \\ &= i_{ts} + e''_{ts}, \quad e''_{ts} \rightarrow 0 \text{ as } \dot{\theta} \rightarrow \theta_0 \\ &\rightarrow i_{ts} \text{ as } \dot{\theta} \rightarrow \theta_0. \end{aligned} \quad \dots \quad (3.5.4)$$

We may write the equation (3.5.2) as the equivalence

$$\sqrt{n} z_s \sim \Sigma j_{ts} \sqrt{n} d_t. \quad (3.5.5)$$

By the assumption that $|i_{ts}| \neq 0$, when $\dot{\theta}$ is sufficiently close to θ_0 , $|j_{rs}| \neq 0$ in virtue of (3.5.4). Therefore, the equivalence (3.5.5) may be written

$$\sqrt{n} d_s \sim \Sigma j^{ts} \sqrt{n} z_t, \quad (j^{ts}) = (j_{ts})^{-1} \quad \dots \quad (3.5.6)$$

$$\sim \Sigma i^{ts} \sqrt{n} z_t \text{ as } j^{ts} \rightarrow i^{ts} \text{ in probability} \quad \dots \quad (3.5.7)$$

which shows that the a.d. of $\sqrt{n} d_1, \dots, \sqrt{n} d_a$ is multivariate normal with dispersion matrix Λ^{-1} . Inverting the relation (3.5.7) we have

$$\sqrt{n} z_s \sim \Sigma i_{ts} \sqrt{n} d_t = \sqrt{n} \Sigma i_{ts} d_t$$

which is (iii) of the Assumption 3.

The Assumption 2 used in Lemma 8 is important as it specifies the condition under which we can assert that the m.l. estimate has the properties mentioned in Assumption 3. But if merely the existence of estimates satisfying Assumption 3 has to be established, Assumption 2 may be replaced by the weaker identifiability condition.

Lemma 9 : *Assumptions 1 and 2b imply Assumption 3.*

We consider a sphere of radius δ round θ_0 . Since

$$\inf \Sigma \pi_i(\theta_0) \log \frac{\pi_i(\theta_0)}{\pi_i(\theta)} \quad \dots \quad (3.5.8)$$

over the sphere is attained, because of continuity of $\pi(\theta)$, at some point on the sphere the identifiability Assumption 2b ensures that the expression (3.5.8) is greater than $\epsilon > 0$. The argument of Lemma 8 applied to points over the chosen sphere shows

that the likelihood at θ_0 exceeds the likelihood of any point on the sphere with probability unity as $n \rightarrow \infty$. This shows that a local maximum of the likelihood is attained at some point inside the sphere. The first derivative of log likelihood vanishes at that point. Hence there exists a root of the m.l. equation which is consistent. The rest of the argument is as in Lemma 8.

3.6. *Goodness of fit tests when the Assumption 3 is not satisfied.* If the estimated parameters do not satisfy the Assumption 3 the statistic $\Sigma(O-E)^2/E$ need not have a χ^2 distribution. But in some cases, as when the method of moments is followed one may construct an alternative statistic with its a.d. as χ^2 on the same d.f. as the goodness of fit χ^2 .

Let the parameters be estimated by the method of moments i.e., by equating the observed moments of the grouped distribution to the hypothetical values for the grouped distribution. As the estimating equations are linear in the frequencies, the estimated deviations

$$p_i - \pi_i, \dots, p_k - \pi_k \quad \dots \quad (3.6.1)$$

are subject to as many linear restrictions as there are independent parameters, besides their sum being zero. If the matrix of derivatives of π_i has rank q , the asymptotic dispersion matrix of the deviations will have rank $(k-q-1)$. We need only choose $(k-q-1)$ independent deviations and using the reciprocal of their dispersion matrix construct the quadratic form. This has χ^2 distribution with $(k-q-1)$ d.f. The asymptotic dispersion matrix of the deviations can be easily computed by the usual methods.

We can also test the significance of deviations in any particular cell or deviations in a particular set of cells as in Section 3.3 using the asymptotic variances and covariances of the deviations (3.6.1).

There is another situation where the parameters are estimated by an efficient method utilizing the individual observations and not simply the observed frequencies in certain class intervals. In such a case the statistic $\Sigma(O-E)^2/E$, when used as $\chi^2(k-q-1)$ over estimates significance. The extent to which this happens has been studied by Chernoff and Lehmann (1954). An alternative expression is given for the excess in terms of the difference between estimates obtained in two different ways.

Let us indicate by θ^* , the estimate of θ , and by π^* that of π , obtained from the original observations by an efficient method of estimation such as the m.l. The information matrix for a single observation is denoted by (j_{rs}) , reserving (i_{rs}) for the grouped distribution. Let $y_s = n^{-1}(\partial \log L / \partial \theta_s)$, where $L = f(x_1, \theta) f(x_2, \theta) \dots f(x_n, \theta)$ and $f(x, \theta)$ is differentiable. We make the following assumption regarding the estimate θ^* .

Assumption 6:

$$\text{plim } \sqrt{n}[y_s - j_{s1}(\theta_1^* - \theta_1^0) - \dots - j_{sq}(\theta_q^* - \theta_q^0)] = 0, s = 1, \dots, q.$$

LARGE SAMPLE TEST CRITERIA

Lemma 11 : Under the Assumptions 3 and 6

- i) a.v. $\sqrt{n}(p_r - \pi_r^*) = v. [\sqrt{n}(p_r - \pi_r)] - \text{a.v.} [\sqrt{n}(\pi_r^* - \pi_r)]$
- ii) a.c. $[\sqrt{n}(p_i - \pi_i^*), \sqrt{n}(p_j - \pi_j^*)]$
 $= \text{c.} [\sqrt{n}(p_i - \pi_i), \sqrt{n}(p_j - \pi_j)] - \text{a.c.} [\sqrt{n}(\pi_i^* - \pi_i), \sqrt{n}(\pi_j^* - \pi_j)]$
- iii) $\text{plim } n \Sigma \frac{(p_r - \pi_r)(\pi_r - \pi_r^*)}{\pi_r} = 0.$

The results (i) and (ii) are immediate consequences of the equations

$$\text{a.c.} [\sqrt{n}(p_r - \pi_r^*), \sqrt{n}(\pi_s^* - \pi_s)] = 0 \text{ for all } r \text{ and } s \quad \dots (3.6.2)$$

which are true if

$$\text{a.c.} [\sqrt{n}(p_r - \pi_r^*), \sqrt{n} y_s] = 0 \text{ for all } r \text{ and } s. \quad \dots (3.6.3)$$

Let us assume for simplicity that the ranks of (j_{rs}) and (i_{rs}) are both equal to q , the number of parameters. To prove (3.6.3) we have to show that

$$\begin{aligned} \text{a.c.} [\sqrt{n}(p_r - \pi_r), \sqrt{n} y_s] &= \text{a.c.} [\sqrt{n}(\pi_r^* - \pi_r), \sqrt{n} y_s]. \\ &= \frac{\partial \pi_r}{\partial \theta_s} \text{ (obtained by expanding } \pi_r^* \text{ in terms of } y_i) \quad \dots (3.6.4) \end{aligned}$$

Since $E(p_r) = \pi_r$, i.e.,

$$\int p_r f(x_1) \dots f(x_n) dx_1 \dots dx_n = \pi_r$$

differentiating with respect to θ_s , we obtain

$$\int p_r L \frac{\partial \log L}{\partial \theta_s} dx_1 \dots dx_n = \int (\sqrt{n} p_r) (\sqrt{n} y_s) L dx_1 \dots dx_n = \frac{\partial \pi_r}{\partial \theta_s} \quad \dots (3.6.5)$$

Comparing (3.6.4) and (3.6.5) we find (3.6.3) is true. Further $\sqrt{n}(\pi_r - \pi_r^*) \sim \sqrt{n} \Sigma \frac{\partial \pi_r}{\partial \theta_s} (\theta_s - \theta_s^*)$. Hence result (iii) of the lemma follows if

$$\text{plim } \sqrt{n} \Sigma \frac{\partial \pi_r}{\partial \theta_s} \frac{(p_r - \pi_r)}{\pi_r} = 0$$

which is true in virtue of (3.4.3).

The results of Lemma 11 are important in many ways. For instance the significance of the deviations in any given set of cells can be tested, although the estimates are not obtained from the grouped distribution, since the variances and covariances of the deviations can be easily computed by using the results (i) and (ii). Thus

$$\begin{aligned} \text{a.v.} [\sqrt{n}(p_r - \pi_r^*)] &= \pi_r(1 - \pi_r) - \Sigma \Sigma \frac{\partial \pi_r}{\partial \theta_s} \frac{\partial \pi_r}{\partial \theta_t} j^{st} \\ \text{a.c.} [\sqrt{n}(p_r - \pi_r^*), \sqrt{n}(p_u - \pi_u^*)] &= -\pi_r \pi_u - \Sigma \Sigma \frac{\partial \pi_r}{\partial \theta_s} \frac{\partial \pi_u}{\partial \theta_t} j^{st}. \end{aligned}$$

It may be observed that the dispersion matrix of all the deviations $(p_i - \pi_i^*)$ may have the maximum rank $(k-1)$, so that a χ^2 based on $(k-1)$ degrees of freedom may be

constructed to test goodness of fit. But this is not likely to be an efficient test. On the other hand the statistic

$$n \sum \frac{(p_i - \pi_i^*)^2}{\pi_i^*}$$

need not have a χ^2 distribution asymptotically as shown below. We use the result (iii) of Lemma 11 to obtain the decomposition

$$n \sum \frac{(p_i - \pi_i^*)^2}{\pi_i^*} \sim n \sum \frac{(p_i - \pi_i)^2}{\pi_i} + n \sum \frac{(\pi_i - \pi_i^*)^2}{\pi_i^*}.$$

The first term on the right hand side is asymptotically $\chi^2(k-q-1)$ as shown in Theorem 1. The second term which is asymptotically independent of the first provides the excess and depends on the numerical difference between estimates obtained in two ways.

4. TESTS IN CONTINGENCY TABLES

Problems associated with a single multinomial distribution have been fully discussed in the earlier sections. Another group of problems, which is important in practice, is related to contingency tables, or independent samples from a number of multinomial distributions. Let us suppose that we have samples of sizes $n_1, \dots, n_m$ ($n_1 + \dots + n_m = n$) from m finite multinomial populations, all not necessarily having an equal number of cells. One of the problems is to test the hypothesis that the cell probabilities could be represented in terms of q parameters, $\theta_1, \dots, \theta_q$. The likelihood of the parameters is the product of the likelihoods

$$L_1 L_2 \dots L_m$$

corresponding to the m samples. Let us define

$$z_{ij} = \frac{1}{n_i} \frac{\partial \log L_i}{\partial \theta_j}$$

$$n_i = n\lambda$$

and (i_{rs}^t) is the information matrix per single observation from the t -th multinomial distribution. We make the following assumptions.

Assumption 7: The rank of (g_{rs}) is q , where $g_{rs} = \sum \lambda_i i_{rs}^t$.

Assumption 8: Every cell probability π_{ij} as a function of the parameters admits continuous first order differential coefficients.

Assumption 9: If $\pi_{i1}, \dots, \pi_{ik_i}$ are the probabilities for the i -th multinomial then

$$\sum_{j=1}^{k_i} \pi_{ij}(\theta_0) \log \frac{\pi_{ij}(\theta_0)}{\pi_{ij}(\theta)}$$

is bounded away from zero when $(\theta - \theta_0)^2 \geq \delta$, however small, for each i .

LARGE SAMPLE TEST CRITERIA

Following the arguments of Lemma 8 it can be shown that there exist consistent estimates $\dot{\theta}_1, \dots, \dot{\theta}_q$ having asymptotic distributions, such that

$$\text{plim } \sqrt{n}[\sum \lambda_i z_{ij} - g_{j1}(\dot{\theta}_1 - \theta_1^0) - \dots - g_{jq}(\dot{\theta}_q - \theta_q^0)] = 0$$

for $j = 1, \dots, m$ and fixed $\lambda_1, \dots, \lambda_m$. We may now prove the following in the same way as Theorems 1 and 2.

Theorem 3: *Under the Assumptions 7, 8 and 9 the asymptotic distribution of the statistic*

$$\sum_{i=1}^m \sum_{j=1}^{k_i} \frac{n_i(p_{ij} - \pi_{ij})^2}{\pi_{ij}} = \sum \sum \frac{(O - E)^2}{E}$$

is $\chi^2(\sum k_i - m - q)$.

We use the decomposition

$$\sum \sum \frac{n_i(p_{ij} - \pi_{ij})^2}{\pi_{ij}} \sim \sum \sum \frac{n_i(p_{ij} - \pi_{ij})^2}{\pi_{ij}} + \sum \sum \frac{n_i(\pi_{ij} - \pi_{ij})^2}{\pi_{ij}} \quad \dots \quad (4.1)$$

where the left hand side is asymptotically $\chi^2(\sum k_i - m)$ and the last term in (4.1) is asymptotically $\chi^2(q)$.

Theorem 4: *Under the Assumptions 7, 8 and 9 and the further assumptions that (i) $\theta_1, \dots, \theta_q$ can be represented as functions of $t < q$ parameters $\alpha_1, \dots, \alpha_t$, (ii) every θ_i admits first partial derivatives in α_j , which are continuous and (iii) a condition similar to Assumption 9 in terms of α_i is satisfied, the a.d. of*

$$\sum_i \sum_j \frac{n_i(\pi_{ij} - \pi_{ij}^*)^2}{\pi_{ij}^*} = \sum \sum \frac{(E_1 - E_2)^2}{E_2}$$

is $\chi^2(q - t)$, where π_{ij}^* denotes the estimate of π_{ij} as a function of parameters α_i , and E_1, E_2 are the expectations under the original and new specifications.

To prove the result we first obtain the decomposition

$$\sum_i \sum_j \frac{n_i(\pi_{ij} - \pi_{ij})^2}{\pi_{ij}} \sim \sum \sum \frac{n_i(\pi_{ij} - \pi_{ij}^*)^2}{\pi_{ij}^*} + \sum \sum \frac{n_i(\pi_{ij}^* - \pi_{ij})^2}{\pi_{ij}}$$

or

$$\sum \sum \frac{(E_1 - H)^2}{H} \sim \sum \sum \frac{(E_1 - E_2)^2}{E_2} + \sum \sum \frac{(E_2 - H)^2}{H}$$

and proceed as in Theorem 2.

Theorems 3 and 4 concerning several multinomial distributions cover the multidimensional contingency tables in so far as hypotheses specifying the cell probabilities are concerned.

As an application let us consider the phenotypic frequencies of O, A, B, AB , blood groups in two samples of individuals from two communities. The first hypothesis is that for each community the frequencies are consistent with Bernstein's theory. There are then four parameters, p_1, q_1 , representing the A and B gene frequencies in one community and p_2, q_2 , in the second. The second hypothesis specifies further that the gene frequencies in the two communities are same, so that all the cell

probabilities involve only two parameters p, q . For estimating these parameters by the m.l. method reference may be made to Rao (1952).

phenotype	community 1			community 2		
	O	E_1	E_2	O	E_1	E_2
O	121	118.89	117.94	118	119.93	121.62
A	120	122.44	105.52	95	92.68	108.81
B	79	81.54	98.14	121	118.74	101.19
AB	33	30.13	31.40	30	32.65	32.37

The test for the first specification (consistency with Bernstein's theory) is

$$\sum \frac{(O-E_1)^2}{E_1} = 0.44 \quad \text{for community 1, d.f.} = 1$$

$$= 0.35 \quad \text{for ,, 2, d.f.} = 1$$

or a total of 0.79, which is small for $\chi^2(2)$. To test the equality of gene frequencies the statistic is

$$\sum \sum \frac{(E_1-E_2)^2}{E_2} = 11.04, \text{ with } 4-2 = 2 \text{ d.f.}$$

This is significant at 1% level, indicating differences in the gene frequencies. In a previous paper (Rao, 1948), the author examined the second hypothesis by using a different large sample test. The present test based on the difference between expected values under the two hypotheses seems to be more attractive in practice.

REFERENCES

- CHERNOFF, H. and LEHMANN, E. L. (1954): The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Ann. Math. Stat.*, **25**, 579-586.
- COCHRAN, W. G. (1954): Some methods of strengthening χ^2 tests. *Biometrics*, **10**, 417.
- CRAMÉR, H. (1946) *Mathematical Methods of Statistics*, Princeton University Press.
- FISHER, R. A. (1921): On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc., A*, **222**, 309.
- (1925) Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, **22**, 700.
- HALDANE, J. B. S. (1951): A class of efficient estimates of a parameter. *Bull. Int. Stat. Inst.*, **33**, 231.
- MITRA, S. K. (1955): Statistical analysis of categorical data. No. 142, Mimeographed Series, Institute of Statistics, North Carolina.
- NEYMAN, J. (1949): Contributions to the theory of the χ^2 test. First Berkeley Symposium on Mathematical Statistics and Probability, 239.
- RAO, C. R. (1948): Large sample tests of statistical hypotheses concerning several parameters with application to problems of estimation. *Proc. Camb. Phil. Soc.*, **44**, 50.
- (1955a): Theory of the method of estimation by minimum chisquare. Proc. 29th Session of the International Statistical Institute, 25-32.
- (1955b): Analysis of dispersion for multiply classified data with unequal numbers in cells. *Sankhyā*, **15**, 253.
- (1958): Maximum likelihood estimation for the multinomial distribution. *Sankhyā*, **18**, 139.
- (1960a): Apparent anomalies and irregularities in maximum likelihood estimation. 32nd session of the International Statistical Institute, Tokyo, 1960.
- (1960b): Asymptotic efficiency and limiting information (to appear in the Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability).
- RAO, C. R. and CHAKRAVARTY, I. M. (1956): Some small sample tests of significance for a Poisson distribution. *Biometrics*, **12**, 264-282.
- WALD, A. and WOLFOWITZ, J. (1944): Statistical tests based on permutations of the observations. *Ann. Math. Stat.*, **15**, 358-372.

Paper received : November, 1960.

A METHOD OF FRACTILE GRAPHICAL ANALYSIS*

By P. C. MAHALANOBIS

Indian Statistical Institute

Fractile Graphical Analysis is a new method of statistical analysis which provides an effective summary of information particularly useful in situations where the data do not permit a description in terms of a few parameters relating to the distribution; it also provides a graphical way of testing differences between groups. This method can be used for any variate which can be ranked. In this paper, the use of this method is illustrated for the comparison of economic data—relating either to the same population at different points of time or to different populations—by means of examples taken from the Indian National Sample Survey.

This paper provides some examples of the use of fractile graphical analysis, a new method for the comparison of economic data relating to the same population over time or to any two populations that differ as to geographical region or in any other way. This method can be used for any variate that can be ranked, and is based on certain theoretical conjectures. Asymptotic, but not exact proofs are available for some of these conjectures and results of model sampling experiments have been found to be in accordance with them.

1. THE METHOD

1.1. In the National Sample Survey of India much economic and demographic data are collected every year in the successive "rounds" of a survey, each "round" extending over several months. The data are usually tabulated separately for the different States (which constitute the Union of India) or for groups of States, for India as a whole and with breakdowns, or in some cases, for rural and urban areas. Consider, for example, surveys of household budgets. The total or per capita consumption expenditure for 30 days (or any given reference period) and also the per capita expenditure on, say, foodgrains, all items of food, drugs, or cloth, etc., would be reported for each sample household. As a probability sample is used in each case, it is possible to estimate any of the characteristics for the whole population. In this way, information is available on the distribution of households by size of total or per capita consumption expenditure, or of the expenditure on individual items, for each size class, for the different States and for India, and over time from round to round of the National Sample Survey. The question naturally arises whether the pattern of consumption is or is not the same from one State to another; is or is not steady over time from one round of the survey to another, for the same State or for the whole of India.

* This paper originally appeared in *Econometrica*, Vol. 28, 2 (April 1960) and is being reprinted with the permission of the Editor, *Econometrica*.

1.2. The design of the survey is always such that Inter-Penetrating Samples (IPS) are used in which two or more independent samples are drawn in exactly the same way—according to the design of the survey and with replacement. These two or more interpenetrating samples are statistically equivalent and supply independent and equally valid estimates of all population characteristics. For convenience these two or more samples may be called sub-samples, emphasizing the fact that they can be pooled together to give one combined sample to supply valid estimates of population characteristics based on all of the information.

1.3. Consider a sample from a given bivariate population, and let x represent the per capita consumption expenditure and y the per capita expenditure on foodgrains for a sample household. Assume that the size of each sub-sample is the same and consists of N households each.¹ Also assume that each sample household has the same chance of being included in each sub-sample. Then all sample households would have the same probability weight for purposes of estimating the population characteristics. The processing of the data in this case is extremely simple. Consider the first sub-sample of N households; rank the households in ascending order of the x -variate (in this case, the per capita expenditure of the household in the given reference period).² Now divide the N sample units into g equal groups of n each, so that $N = gn$. Next, find the average value of the y variate, for example, the per capita expenditure on foodgrains for each of the g groups. These may be called $y_{11}, y_{12}, y_{13}, \dots, y_{1g}$. Now take g equidistant points³ on the x -axis and plot the y values, that is, $y_{11}, y_{12}, y_{13}, \dots, y_{1g}$ on these g points $x = 1, 2, \dots, g$. Finally, join the successive points by straight lines.⁴ This would give a graph which may be called $G(1)$ for the first sub-sample.

1.4. The procedure would be slightly more complicated when the sample units are selected with varying probabilities of being included in the sample, and, therefore, have varying probability weights for purposes of estimating the population characteristics. Consider the case in which x_i is the per capita total expenditure and y_i is the per capita expenditure on foodgrains for the i -th sample household. Let w_i be the probability weight of this household for estimating the population characteristics; also, let the w_i 's be adjusted in such a way that the total probability weight, that is, the sum of all values of w_i , equals one. The first stage of processing, namely, ranking the sample households in ascending values of x_i , remains the same; the next stage is to multiply y_i (and other similar variates if there are more than one) by w_i ;

¹ This can be arranged by specification in the design of the survey, but this is not essential. One great advantage of the fractile graphical method is that the procedure would remain the same even when the sub-samples are of different sizes.

² This can be done by hand or very quickly with punched cards by a sorter.

³ These points would be, of course, centred at the midpoint of each of the g groups. It is, however, not necessary that the x -points should be equidistant; other scales can be used, for example, values of x such that the expected frequency in each group would be equal for a normal distribution. Any other suitable distribution may also be used.

⁴ It is possible to use other rules of construction (for example, graduating parabolas passing through three or more successive points) to draw the graph.

A METHOD OF FRACTILE GRAPHICAL ANALYSIS

and then to form the accumulated sums of both w_i and $w_i y_i$ for all sample households.⁵ It is now possible to form any desired number of fractile groups, say, g , each having an equal proportion of the estimated total population of households. For example, if $g = 20$, all that has to be done is to divide the sub-sample at successive five per cent points into groups on the basis of the accumulated sums of w_i . It is also easy to obtain the estimated average value of any population characteristic y for each of the fractile groups from the corresponding accumulated sums of $w_i y_i$. These average values of y are then plotted against g equidistant points on the x -axis, and successive plotted points are joined by straight lines to supply the graph $G(1)$.

1.5. Consider next the second sub-sample, also of N sample households. Rank them by ascending values of x ; divide into g equal fractile groups; find the average values $y_{21}, y_{22}, \dots, y_{2g}$ of the y variate for each of the g groups; plot these points on the same chart and on the same g equidistant points as used for the first graph, $G(1)$; and join the successive points by straight lines. This would give a second graph, $G(2)$, for the second sub-sample.

1.6. The next step is to pool the two sub-samples together to form a single sample of size $2N$. Rank the sample units again by ascending values of x ; group them into g equal groups each of size $2n$; find the average values of y , that is, $y_1, y_2, \dots, y_g$ for each of the g groups; plot them on the same chart and on the same g equidistant points; and join the successive points by straight lines. This would give the combined graph which may be called $G(1, 2)$ for the combined sample. This completes the construction for the given population.⁶

1.7. I shall now make some assumptions which seem plausible. Consider the area bounded by the two sub-sample graphs $G(1)$ and $G(2)$. (This can be lightly shaded in the chart to make it distinct). We may use this area as a convenient measure of "error"⁷ to be associated with the combined graph $G(1, 2)$.

⁵ If punched cards are used, this can easily be done by running them through a tabulator and printing the accumulated subtotals for the desired fractile groups, or for fixed ranges of the x variate. A convenient plan is to use twenty 5 per cent groups together with five 1 per cent groups at both the bottom and the top, giving thirty groups altogether.

⁶ A complete or exhaustive processing would be given if the accumulated totals of wy were obtained for each successive sample household. This can easily be done by running the punched cards through a tabulator and printing all successive accumulated totals. The three sets of accumulated sums contain the whole of the information. The three graphs $G(1)$, $G(2)$, and $G(1, 2)$, would also contain the information in a form such that the accuracy with which the information could be recovered would depend on the accuracy with which observations could be read from the graphs. Also, it may be possible to graduate the three sets of accumulated sums by suitable mathematical functions; and the information can be recovered from such functions with a margin of uncertainty which would depend on the "goodness of fit" of the graduating functions and the margins of error given by the sub-samples.

Let $F(1)$, $F(2)$ and $F(1, 2)$ be the three graduating functions of the same specification which are fitted to the data for the two sub-samples and for the combined sample, respectively. These three curves can be easily plotted on the chart. Then the area bounded by the two graduating functions for the two sub-samples $F(1)$ and $F(2)$ would supply a measure of the error to be associated with the graduating function for the combined sample. Alternatively, the deviation from the graduating function can be measured directly by the area bounded by the observed combined graph $G(1, 2)$ and the corresponding graduating curve $F(1, 2)$. These two areas should be statistically of the same order of magnitude; and if these areas (that is, the deviation from the graduating function) are less than the margin of error given by the area bounded by the two sub-samples, it would then be possible to recover from the graduating function almost the whole of whatever information can validly be used for statistical purposes relating to the estimated population characteristics (although some of the sample-specific information may be lost).

⁷ This and other conjectures are given in an Appendix.

1.8. Now consider a second population and assume that a pair of interpenetrating samples have been taken from it with the same sample size of N households in each sub-sample. It is then possible to go through the same construction as in the case of the pair of sub-samples from the first population. Let the corresponding graphs for the second population be called $G'(1)$, $G'(2)$ and $G'(1, 2)$. The area between the sub-sample graphs $G'(1)$ and $G'(2)$ would then give the corresponding "error"⁸ associated with the combined graph $G'(1, 2)$ for the second population.

1.9. It is now possible to go one step further. The area between $G(1, 2)$ and $G'(1, 2)$ may be used as a measure of the "separation," or over-all difference or generalised distance, between the two populations. It is plausible to assume that the "error" to be associated with the "separation" can be derived in the usual way⁹ from the two component errors associated respectively with each of the combined graphs $G(1, 2)$ and $G'(1, 2)$. It is also possible to consider not only the total separation but the separation between any particular portions of the two combined graphs $G(1, 2)$ and $G'(1, 2)$, that is, the area bounded by these two graphs and the corresponding ordinates limiting any assigned fractile group. Each such partial separation would also have its associated error which can be derived from the two component errors lying between the same two ordinates. This furnishes a convenient tool for the comparison, analysis, and the testing of significance of the separation between two populations either over the whole or any part of the range of observations.

2. ILLUSTRATIONS FROM THE NATIONAL SAMPLE SURVEY OF INDIA

2.1. Some actual examples from the National Sample Survey of India may now be considered. In the 7th round (October, 1953 to March, 1954) information on household consumption was collected in the form of two interpenetrating samples of 702 and 711 sample households from 476 and 478 *mauzas* or villages, respectively, extending over the whole of rural India (excluding Jammu and Kashmir). In another survey carried out in the 9th round (May to November, 1955) similar information was collected from 768 households, one each from 768 *mauzas* (out of 772 *mauzas* or villages of which four were uninhabited)¹⁰ in each of the two sub-samples. The period of reference was 30 days in each of the surveys. It is possible to use the per capita expenditure on all items of consumption as the x variate. The sample households were ranked (separately for each of the two sub samples and for the combined sample) in ascending order of the x variate (per capita consumption expenditure). The design of the sample surveys was with varying probabilities, and appropriate probability

⁸ Although sub-samples have been assumed to be of equal sizes, this is not necessary. It is also unnecessary that the size of the sample should be the same in the case of the two samples from the two populations. Let N_1 , N_2 be the sizes of the two sub-samples from the first population and N'_1 and N'_2 be the sizes of the two sub-samples from the second population. Then whether $N_1 = N_2 = N'_1 = N'_2$ (as assumed above), or whether they are all different is entirely immaterial. The graphical measure of the two "errors" would be given in every case by the two areas lying respectively between each pair of sub-sample graphs $G(1)$ and $G(2)$ or $G'(1)$ and $G'(2)$.

⁹ That is, by the square root of the sum of the squares of the errors of the two combined graphs $G(1, 2)$ and $G'(1, 2)$.

¹⁰ The word "village" is used for the Indian revenue unit *mauza* which broadly corresponds to a village. Although these revenue units are several centuries old and some have become uninhabited, land continues to be demarcated in terms of the old *mauza* units.

A METHOD OF FRACTILE GRAPHICAL ANALYSIS

weights were used for estimating the population characteristics. The estimated number of households and other characteristics of the population were obtained in the usual way, and twenty equal fractile groups, each containing 5 per cent of the households, were formed. The bottom and top 5 per cent groups were subdivided into 1 per cent groups (1, 2, 3, 4, 5 and 96, 97, 98, 99, 100 percentiles), giving 30 groups altogether. Some additional fractile groups were also formed.¹¹

2.2. Tables I and II give the information for the 7th and 9th rounds respectively.¹² In each table, columns (2), (3) and (4) give the limiting value of the per capita expenditure at the upper end of the fractile group for (respectively) sub-sample 1 (which would be written in the contracted form s.s. 1), sub-sample 2 (s.s. 2) and the combined sample. For example, in Table I for the bottom 1 per cent group, the limiting values of per capita expenditure are Rs. 2.60 and Rs. 3.25 in the two sub-samples and Rs. 3.00 in the combined sample. Columns (5), (6), and (7) in both the tables give the average per capita consumer expenditure in each fractile group. It is known that prices of consumer goods, especially of foodgrains, had greatly decreased at the time of the 9th round. The effect of this fall in prices can be seen from these two tables. Consider the fractile group between 70 per cent and 75 per cent of households; the lower and upper limiting values of per capita expenditure in the 7th round are, respectively, Rs. 20.00 and Rs. 22.00. It should be noticed, however, that the fractile group between 75 per cent and 80 per cent in the 9th round has approximately the same limiting values—Rs. 20.02 and Rs. 22.11. If, in terms of money values, a fixed range had been used, the group having a per capita expenditure between Rs. 20 and Rs. 22 per 30 days, would have been between 70 and 75 per cent of the households in the 7th round, but would have been almost between 75 and 80 per cent in the 9th round. In the present approach the comparison between the two rounds is made on the basis of the *same* fractile group, that is between 70 and 75 per cent or between 75 and 80 per cent of households in each of the rounds.

2.3. Consider also an associated variate y , the per capita expenditure on foodgrains for each sample household. The average value of y , that is, the average per capita expenditure on foodgrains for each fractile group, is given in columns (2), (3), and (4) of Table III for the first sub-sample (s.s. 1), the second sub-sample (s.s. 2) and the combined sample for the 7th round respectively. Corresponding values for the 9th round are given in columns (2), (3), and (4) of Table IV. The results are shown in Figure 1, where the x -axis represents percentages of households, as ranked by per capita consumption expenditure. The limiting values of the per capita expenditure at the upper end of the each fractile group (in Table I, for the 7th round) are shown at the top of the x -scale; and the corresponding

¹¹ Averages for larger fractile groups can be obtained directly by taking the average of an appropriate number of equi-frequency fractile groups.

¹² Based on the sample survey, the estimated number of rural households was 63.4 million in the 7th round and 65.3 million in the 9th round; the estimated rural population was 324 million and 338 million in the 7th and 9th rounds, respectively; and the estimated per capita consumption expenditure in rural households was Rs. 5565 million and Rs. 5131 million per 30 days at current prices in the 7th and the 9th rounds, respectively. The total number of *mauzas* (villages) in India, excluding Jammu and Kashmir, was 603,168 in the 1951 census.

SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS : SERIES A

TABLE I. NATIONAL SAMPLE SURVEY OF INDIA: ALL INDIA, RURAL
7TH ROUND: OCTOBER, 1954—MARCH, 1955*
TOTAL PER CAPITA CONSUMER EXPENDITURE IN RUPEES PER 30 DAYS

serial num- ber	fractile group (percent)	total per capita consumer expenditure in rupees per 30 days					
		limiting value at upper end of each fractile group			average expenditure in each fractile group		
		s.s.1	s.s.2	combined	s.s.1	s.s.2	combined
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	1	2.60	3.25	3.00	2.44	2.20	2.23
2	2	3.00	4.13	3.38	2.91	3.74	3.14
3	3	4.00	4.75	4.57	3.35	4.54	4.10
4	4	5.00	4.83	4.83	4.63	4.78	4.73
5	5	5.57	5.25	5.50	5.32	4.99	5.13
6	0—5	5.57	5.25	5.50	3.75	3.99	3.85
7	5—10	7.25	6.88	7.11	6.63	6.24	6.39
8	10—15	8.25	7.90	8.00	7.72	7.46	7.54
9	15—20	9.29	9.00	9.25	8.77	8.53	8.64
10	20—25	9.80	9.67	9.75	9.58	9.43	9.51
11	25—30	10.60	10.25	10.50	10.22	9.93	10.03
12	30—35	11.83	11.00	11.40	11.24	10.69	10.95
13	35—40	13.00	12.00	12.43	12.44	11.45	11.94
14	40—45	14.00	13.00	13.50	13.37	12.48	12.83
15	45—50	14.75	14.25	14.40	14.23	13.64	14.00
16	50—55	16.00	15.25	15.67	15.40	14.73	15.02
17	55—60	17.60	16.60	17.00	16.82	15.99	16.37
18	60—65	18.80	18.00	18.45	18.34	17.22	17.78
19	65—70	21.00	19.25	20.00	19.90	18.67	19.18
20	70—75	22.00	22.00	22.00	21.39	20.49	21.11
21	75—80	25.00	24.00	24.33	23.34	23.02	23.12
22	80—85	27.67	28.00	28.00	26.63	26.09	26.40
23	85—90	32.75	32.33	32.67	30.07	29.74	29.95
24	90—95	43.67	39.00	41.33	37.17	35.81	36.14
25	95—100	226.00	264.50	264.50	74.11	58.97	64.98
26	96	47.00	43.00	45.33	45.76	41.97	43.86
27	97	59.00	48.00	51.00	51.91	45.51	47.34
28	98	72.17	53.50	61.00	66.35	51.37	54.08
29	99	76.00	65.25	74.33	75.46	59.64	67.05
30	100	226.00	264.50	264.50	118.11	142.48	122.59
31	0—20				6.74	6.49	6.64
32	20—40				10.88	10.49	10.66
33	40—60				14.92	13.97	14.44
34	60—80				20.52	19.91	20.20
35	80—100				39.77	36.45	38.03
36	0—25				7.30	7.11	7.18
37	25—50				12.41	11.73	12.03
38	50—75				18.20	17.45	17.83
39	75—100				36.45	33.44	34.86
40	0—50				9.88	9.58	9.73
41	50—100				26.57	25.20	25.87
42	0—100				17.65	16.28	17.24

	sub-sample 1	sub-sample 2	combined
* number of sample villages	476	478	954
number of sample households	702	711	1413

A METHOD OF FRACTILE GRAPHICAL ANALYSIS

TABLE II. NATIONAL SAMPLE SURVEY OF INDIA : ALL INDIA, RURAL
9TH ROUND : MAY—NOVEMBER, 1955*

TOTAL PER CAPITA CONSUMER EXPENDITURE IN RUPEES PER 30 DAYS

serial num- ber	fractile group (percent)	total per capita consumer expenditure in rupees per 30 days					
		limiting value at upper end of each fractile group			average expenditure in each fractile group		
		s.s.1	s.s.2	combined	s.s.1	s.s.2	combined
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	1	3.19	2.98	3.16	2.99	2.58	2.71
2	2	3.91	3.71	3.87	3.67	3.43	3.52
3	3	4.28	4.16	4.20	4.13	4.01	4.06
4	4	4.65	4.56	4.65	4.51	4.39	4.46
5	5	4.91	4.79	4.90	4.85	4.72	4.78
6	0—5	4.91	4.79	4.90	4.01	3.82	3.90
7	5—10	5.99	5.85	5.92	5.43	5.39	5.39
8	10—15	6.74	6.88	6.80	6.39	6.38	6.38
9	15—20	7.54	7.90	7.76	7.25	7.35	7.28
10	20—25	8.52	8.59	8.56	8.05	8.31	8.20
11	25—30	9.36	9.57	9.53	8.96	9.19	9.06
12	30—35	10.17	10.39	10.31	9.80	9.95	9.89
13	35—40	11.03	11.09	11.07	10.66	10.74	10.69
14	40—45	11.80	12.06	11.92	11.42	11.68	11.49
15	45—50	12.83	13.02	12.88	12.35	12.59	12.43
16	50—55	14.14	14.05	14.14	13.46	13.45	13.44
17	55—60	15.13	15.55	15.24	14.57	14.73	14.64
18	60—65	16.61	16.97	16.77	15.81	16.12	16.00
19	65—70	18.49	18.59	18.56	17.61	17.88	17.78
20	70—75	19.85	20.08	20.02	19.14	19.37	19.26
21	75—80	22.59	22.01	22.11	21.22	21.27	21.19
22	80—85	24.95	23.60	24.35	23.77	22.97	23.36
23	85—90	29.57	27.53	28.59	27.40	25.42	26.23
24	90—95	38.05	38.38	38.13	32.85	32.06	32.17
25	95—100	194.41	128.86	194.41	55.96	53.30	54.31
26	96	43.45	40.74	40.77	40.34	39.83	39.35
27	97	47.41	44.16	46.30	45.60	42.94	44.01
28	98	58.18	57.16	57.16	56.06	49.80	52.11
29	99	66.86	67.69	66.86	65.34	72.22	62.86
30	100	194.41	128.86	194.41	96.54	66.65	78.88
31	0—20				5.64	5.72	5.67
32	20—40				9.39	9.56	9.46
33	40—60				13.01	13.15	13.06
34	60—80				18.33	18.82	18.58
35	80—100				32.90	33.15	32.99
36	0—25				6.06	6.26	6.16
37	25—50				10.61	10.77	10.68
38	50—75				16.04	16.27	16.12
39	75—100				30.53	30.45	30.43
40	0—50				8.31	8.48	8.39
41	50—100				22.65	23.10	22.85
42	0—100				15.03	15.24	15.15

* The number of sample villages was 772 (including 4 uninhabited mauzas) and the number of sample households was 768 in each sub-sample.

SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS : SERIES A

TABLE III. NATIONAL SAMPLE SURVEY OF INDIA: ALL INDIA, RURAL 7TH ROUND: OCTOBER, 1953—MARCH, 1954*

AVERAGE PER CAPITA EXPENDITURE ON FOODGRAINS IN RUPEES AND AS
FRACTION OF TOTAL EXPENDITURE PER 30 DAYS

serial num- ber	fractile group (percent)	average per capita expendi- ture on foodgrains			average fraction of total expenditure		
		s.s.1	s.s.2	combined	s.s.1	s.s.2	combined
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	1	1.09	2.83	2.16	0.436	0.588	0.572
2	2	2.32	2.75	2.13	0.751	0.704	0.679
3	3	2.26	2.57	2.74	0.656	0.550	0.656
4	4	3.12	2.66	2.71	0.639	0.624	0.576
5	5	3.70	3.22	3.39	0.632	0.622	0.610
6	0—5	2.49	2.79	2.61	0.635	0.614	0.620
7	5—10	4.07	4.20	4.20	0.600	0.664	0.648
8	10—15	5.07	4.44	4.70	0.616	0.554	0.573
9	15—20	4.99	5.28	5.11	0.562	0.584	0.576
10	20—25	5.71	4.66	5.17	0.572	0.535	0.550
11	25—30	5.56	5.39	5.53	0.560	0.560	0.562
12	30—35	5.59	5.57	5.98	0.515	0.557	0.541
13	35—40	6.00	6.73	6.28	0.499	0.552	0.534
14	40—45	6.75	7.29	6.82	0.487	0.546	0.492
15	45—50	8.15	6.60	7.46	0.510	0.490	0.503
16	50—55	7.38	6.83	7.21	0.481	0.475	0.491
17	55—60	7.67	7.84	7.78	0.492	0.516	0.489
18	60—65	7.31	8.49	7.85	0.398	0.464	0.444
19	65—70	8.95	7.65	8.23	0.446	0.417	0.425
20	70—75	8.22	9.27	8.74	0.423	0.456	0.440
21	75—80	9.44	8.93	9.24	0.388	0.396	0.398
22	80—85	8.69	8.69	8.59	0.358	0.382	0.366
23	85—90	9.30	8.53	8.96	0.324	0.333	0.331
24	90—95	10.13	8.18	9.29	0.292	0.238	0.272
25	95—100	12.02	12.19	11.78	0.214	0.234	0.219
26	96	13.18	13.09	10.38	0.303	0.322	0.261
27	97	13.72	12.76	15.01	0.253	0.277	0.275
28	98	13.29	12.04	11.51	0.233	0.222	0.222
29	99	10.68	10.06	11.74	0.139	0.189	0.185
30	100	9.27	13.21	10.38	0.081	0.161	0.102
31	0—20	4.19	4.13	4.17	0.601	0.603	0.602
32	20—40	5.72	5.70	5.77	0.534	0.552	0.547
33	40—60	7.54	7.16	7.30	0.492	0.506	0.494
34	60—80	8.40	8.59	8.49	0.412	0.432	0.426
35	80—100	9.87	9.25	9.54	0.296	0.295	0.296
36	0—25	4.49	4.24	4.36	0.615	0.597	0.607
37	25—50	6.51	6.47	6.46	0.524	0.551	0.536
38	50—75	7.85	8.02	7.94	0.432	0.460	0.446
39	75—100	9.79	9.18	9.48	0.268	0.274	0.272
40	0—50	5.51	5.43	5.46	0.558	0.567	0.561
41	50—100	8.74	8.58	8.67	0.329	0.341	0.335
42	0—100	7.01	6.90	6.96	0.398	0.410	0.404

	sub-sample 1	sub-sample 2	combined
* number of sample villages	476	478	954
number of sample households	702	711	1413

A METHOD OF FRACTILE GRAPHICAL ANALYSIS

TABLE IV. NATIONAL SAMPLE SURVEY OF INDIA: ALL INDIA, RURAL
9TH ROUND: MAY—NOVEMBER, 1955*

AVERAGE PER CAPITA EXPENDITURE ON FOODGRAINS IN RUPEES AND AS
FRACTION OF TOTAL EXPENDITURE PER 30 DAYS

serial num- ber	fractile group (percent)	average per capita expendi- ture on foodgrains			average fraction of total expenditure		
		s.s.1	s.s.2	combined	s.s.1	s.s.2	combined
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	1	1.72	1.65	1.66	0.581	0.654	0.638
2	2	1.66	1.86	1.87	0.463	0.567	0.532
3	3	2.46	2.36	2.34	0.597	0.604	0.594
4	4	2.48	3.19	2.89	0.556	0.675	0.602
5	5	3.52	3.17	3.25	0.686	0.672	0.657
6	0—5	2.34	2.49	2.39	0.571	0.634	0.603
7	5—10	3.01	3.14	3.09	0.557	0.581	0.568
8	10—15	3.43	3.54	3.48	0.550	0.550	0.554
9	15—20	3.42	3.88	3.60	0.511	0.541	0.524
10	20—25	4.46	4.37	4.50	0.554	0.516	0.539
11	25—30	4.28	4.50	4.39	0.509	0.502	0.500
12	30—35	5.11	5.06	5.05	0.525	0.485	0.509
13	35—40	5.44	5.43	5.34	0.522	0.528	0.517
14	40—45	5.18	6.02	5.61	0.469	0.500	0.493
15	45—50	6.25	5.70	6.00	0.495	0.484	0.488
16	50—55	5.45	6.09	5.80	0.418	0.456	0.434
17	55—60	7.05	5.99	6.56	0.477	0.417	0.456
18	60—65	6.29	6.09	6.30	0.428	0.423	0.427
19	65—70	5.68	7.03	6.18	0.343	0.391	0.355
20	70—75	7.15	7.23	7.19	0.389	0.374	0.383
21	75—80	7.41	7.61	7.50	0.354	0.360	0.356
22	80—85	7.82	8.49	8.05	0.332	0.378	0.354
23	85—90	8.36	8.09	8.21	0.299	0.330	0.316
24	90—95	9.66	7.21	8.50	0.289	0.290	0.288
25	95—100	10.51	8.97	9.50	0.189	0.193	0.190
26	96	7.96	9.81	9.01	0.225	0.254	0.235
27	97	11.06	6.73	8.68	0.246	0.162	0.222
28	98	9.64	8.72	7.31	0.173	0.217	0.165
29	99	12.01	10.75	11.65	0.165	0.201	0.183
30	100	13.98	9.86	11.52	0.142	0.114	0.140
31	0—20	3.00	3.26	3.11	0.548	0.576	0.562
32	20—40	4.81	4.85	4.81	0.528	0.506	0.516
33	40—60	6.03	5.95	6.01	0.465	0.463	0.467
34	60—80	6.56	7.04	6.79	0.379	0.385	0.380
35	80—100	8.92	8.12	8.52	0.277	0.294	0.286
36	0—25	3.25	3.49	3.38	0.537	0.557	0.549
37	25—50	5.24	5.32	5.26	0.494	0.494	0.493
38	50—75	6.30	6.49	6.39	0.393	0.399	0.396
39	75—100	8.61	8.00	8.30	0.282	0.263	0.273
40	0—50	4.24	4.39	4.31	0.510	0.517	0.514
41	50—100	7.35	7.22	7.29	0.325	0.312	0.319
42	0—100	5.70	5.71	5.70	0.379	0.373	0.376

*The number of sample villages was 772 (including 4 uninhabited mauzas) and the number of sample households was 768 in each sub-sample.

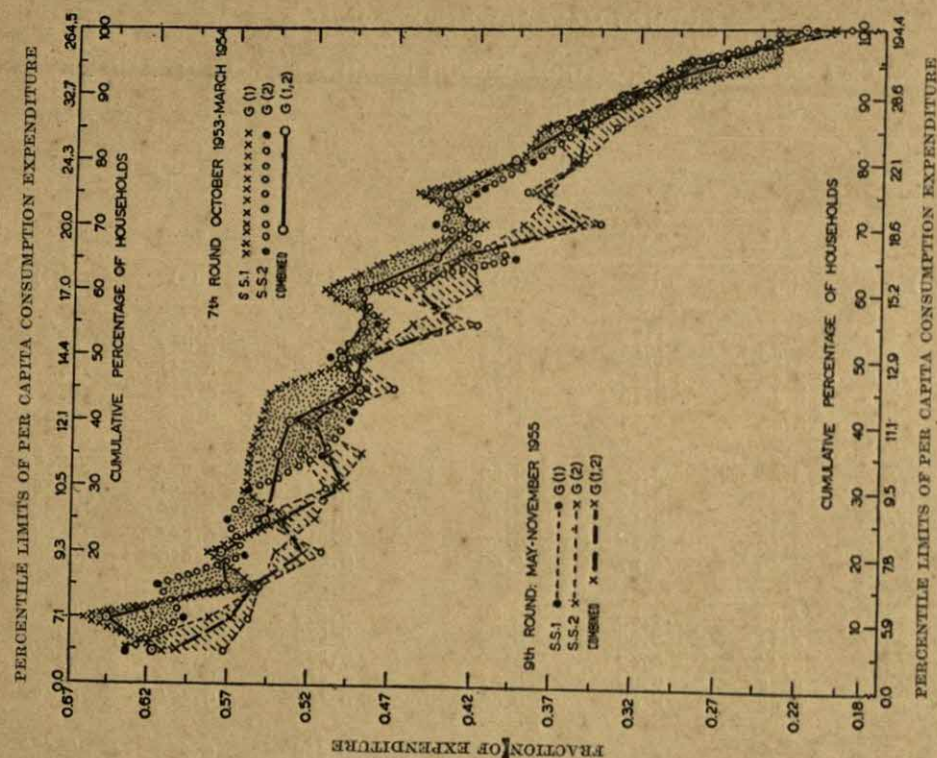


Figure 2. National Sample Survey : all India, rural.
Per capita expenditure on foodgrains as fraction of total expenditure per 30 days

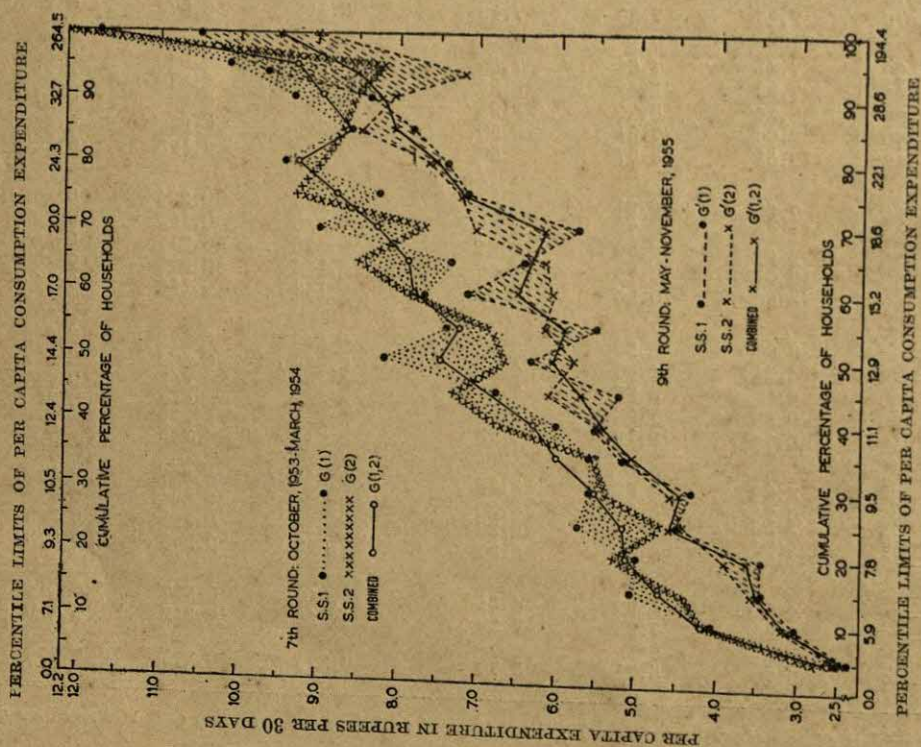


Figure 1. National Sample Survey : all India, rural.
Per capita expenditure on foodgrains per 30 days.

A METHOD OF FRACTILE GRAPHICAL ANALYSIS

limiting values (in Table II for the 9th round) are shown at the bottom of the x -scale. The y -axis represents the average value of the per capita expenditure on foodgrains in rupees per 30 days for each 5 per cent fractile group.

2.4. Now consider the three upper graphs in Figure 1 which represent $G(1)$, $G(2)$ and $G(1, 2)$ for the 7th round. The shaded area between $G(1)$ and $G(2)$ supplies a measure of the error associated with $G(1, 2)$, which has been called e . It will be seen from all three graphs, that the per capita expenditure on foodgrains increases over the whole range from the bottom 5 per cent to the top 5 per cent of households, and that the general trend of the increase in expenditure is significant in comparison with the associated error. The three lower graphs in Figure 1 in the same way represent the three graphs $G'(1)$, $G'(2)$, $G'(1, 2)$ for the 9th round and the shaded area between $G'(1)$ and $G'(2)$ gives the error associated with $G'(1, 2)$ or e' . In this case also the general trend of the increase in expenditure on foodgrains is significant in comparison with the associated error.

2.5. Now consider the "separation" which is given by the area lying between the two combined graphs $G(1, 2)$ and $G'(1, 2)$ for the two rounds of the survey.¹³ The per capita expenditure on foodgrains was lower for every 5 per cent fractile group in the 9th round. From the bottom up to 80 per cent of the households, the separation is seen to be roughly greater than the sum of the two associated error areas,¹⁴ e and e' . The 9th round decrease in the expenditure on foodgrains for up to 80 per cent for households may, therefore, be considered statistically significant. The per capita expenditure on foodgrains was also less in the 9th round for the top 20 per cent of households, but the separation was here smaller than the two associated error areas. It is not possible, therefore, to assert that the decrease was significant for *each* of the four 5 per cent fractile groups at the top. This point can be further examined by pooling together the four 5 per cent fractile groups at the top to form a single 20 per cent group. For this group the average values for the 7th round from Table III are Rs. 9.87 and Rs. 9.25 for the two sub-samples and Rs. 9.54 for the combined sample. For the 9th round from Table IV the average values are Rs. 8.92 and Rs. 8.12 for the two sub-samples and Rs. 8.52 for the combined sample. The difference between the two rounds for the estimates based on the combined samples is Rs. 1.02, which is only somewhat greater than the associated error. Even for the top 20 per cent group as a whole the decrease in the average per capita expenditure on foodgrains was not quite significant:

2.6. Consider next the per capita expenditure on foodgrains as a fraction of the total per capita consumer expenditure of which the average values for each fractile group are given in columns (5), (6), and (7) of Table III and Table IV for the

¹³ The two error areas, e and e' , are seen to be roughly equal, and E , the error of the separation, is therefore very roughly $1.4 e$.

¹⁴ As the sum of the component areas ($e + e'$) is greater than $\sqrt{(e + e')^2}$ there is a margin of safety in using the sum of the errors to test the significance of the separation. The advantage is that a visual comparison is possible. It is also possible to measure the separation, S , as well as the two associated errors e and e' ; calculate E , the error of the separation; obtain S/E and plot these values for each fractile part. This would involve a certain amount of additional calculations, but would supply necessary material in a convenient graphical form to test the significance of the separation.

7th round and the 9th round, respectively. The corresponding graphs are shown in Figure 2. It will be noticed that the values of the fraction is generally lower in the 9th round but the separation is now much less in comparison with the associated error areas. That is, the fraction of total consumer expenditure spent on foodgrains drops to some extent with a fall in prices but, on the whole, tends to remain much more stable than the actual expenditure on foodgrains. For the rural households as a whole, foodgrains account for from 38 to 40 per cent of all consumer expenditure and the fraction is over 60 per cent in the bottom 5 per cent of households.

2.7. The same data can be shown in the form of cumulative percentages which are given for the expenditure on all consumer items in columns (2), (3), and (4) and for the expenditure on foodgrains in columns (5), (6) and (7) in Tables V and VI for the 7th round and 9th round, respectively. Figure 3 shows the graphs for the 7th round, and Figure 4 for the 9th round. The three upper graphs in Figure 3 and Figure 4 represent the concentration of expenditure on foodgrains, and the area lying between the two sub-sample graphs $G(1)$ and $G(2)$ gives the error associated with the graph $G(1, 2)$ based on the combined sample. The three lower graphs represent the concentration of total consumer expenditure. In both cases, the separation between $G(1, 2)$ and $G'(1, 2)$ or between the two sets of three graphs is much greater than either of the two associated error areas. The concentration graph for expenditure on foodgrains thus lies significantly nearer the line of equal distribution, $y = x$, showing the inelastic nature of the expenditure on foodgrains. The area lying between the line $y = x$ and the graph $G(1, 2)$ can supply a convenient measure of concentration; this also will have the same associated error as $G(1, 2)$.

2.8. The concentration curves (based on the combined sample in each case) for both the 7th and 9th rounds are shown in Figure 5. The two lower graphs, which represent the two concentration curves for total consumer expenditure, cross and recross, indicating that there was no change in this respect between the 7th and the 9th round. The two upper graphs represent the two concentration curves for the expenditure on foodgrains. These two graphs also cross and recross up to the 50th percentile of households and again beyond the 90th percentile of households, the graph for the 9th round (when prices were lower) is systematically below the graph for the 7th round. It is possible to examine these portions of the two graphs in greater detail. Figure 6 shows the concentration curves between the 50th and the 90th percentiles of households on a magnified scale for both rounds (with separate graphs for the two sub-samples and the combined sample in each case). The shaded area shows the error associated with the respective graphs. The "separation" of the two graphs $G(1, 2)$ and $G'(1, 2)$ between the two rounds is on the whole somewhat greater than the two associated error areas. This, if real, would indicate that in this middle region of households between the 50th and 90th percentiles (with total per capita consumer expenditure lying roughly between Rs. 13 or 14 and Rs. 30 or Rs. 32 per month) the expenditure on foodgrains tends to behave somewhat more as a necessity when prices are higher. The observed difference is based on very small samples and may have arisen through chance; but it may deserve more careful study with larger samples.

A METHOD OF FRACTILE GRAPHICAL ANALYSIS

TABLE V. NATIONAL SAMPLE SURVEY OF INDIA: ALL INDIA, RURAL
7TH ROUND: OCTOBER, 1952-MARCH, 1954*

CUMULATIVE PERCENTAGE OF TOTAL CONSUMER EXPENDITURE AND EXPENDITURE
ON FOODGRAINS PER 30 DAYS

serial num- ber	fractile group (percent)	cumulative percentage of expenditure			cumulative percentage of expenditure on foodgrains		
		s.s.1	s.s.2	combined	s.s.1	s.s.2	combined
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	1	0.14	0.14	0.13	0.15	0.45	0.31
2	2	0.26	0.36	0.31	0.40	0.85	0.61
3	3	0.50	0.67	0.55	0.79	1.28	1.01
4	4	0.81	0.93	0.86	1.31	1.62	1.45
5	0—5	1.07	1.18	1.11	1.79	2.01	1.86
6	5—10	3.09	3.13	3.00	4.90	5.20	4.93
7	10—15	5.88	5.59	5.54	9.50	8.76	8.85
8	15—20	8.22	7.64	8.01	12.84	11.87	12.47
9	20—25	11.05	10.62	10.71	17.07	15.45	16.10
10	25—30	13.91	13.07	13.35	20.99	18.71	19.70
11	30—35	17.36	16.22	17.44	25.31	22.70	25.23
12	35—40	21.02	21.19	21.05	29.75	29.83	29.93
13	40—45	24.88	26.88	25.68	34.65	37.92	36.02
14	45—50	30.26	30.71	30.47	42.38	42.46	42.34
15	50—55	34.76	34.93	34.69	47.81	47.21	47.35
16	55—60	39.83	39.31	39.53	53.61	52.48	53.05
17	60—65	45.91	44.11	44.64	59.71	58.22	58.63
18	65—70	51.22	49.45	50.50	65.69	63.58	64.86
19	70—75	56.11	55.46	55.75	70.41	70.19	70.23
20	75—80	61.79	62.33	61.98	76.18	76.71	76.39
21	80—85	69.75	70.00	69.99	82.71	82.92	82.84
22	85—90	76.52	77.31	76.94	87.96	88.04	87.99
23	90—95	85.45	87.26	85.84	94.08	93.58	93.65
24	96	87.04	88.19	88.19	95.21	94.29	95.03
25	97	89.37	92.14	90.29	96.74	96.99	96.67
26	98	92.25	94.44	92.64	98.18	98.31	97.91
27	99	93.87	96.83	95.08	98.75	99.28	98.97
28	100	100.00	100.00	100.00	100.00	100.00	100.00

	sub-sample 1	sub-sample 2	combined
* number of sample villages	476	478	954
number of sample households	702	711	1413

SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS : SERIES A

TABLE VI. NATIONAL SAMPLE SURVEY OF INDIA : ALL INDIA, RURAL
9TH ROUND : MAY, 1955—NOVEMBER, 1955*CUMULATIVE PERCENTAGE OF TOTAL CONSUMER EXPENDITURE AND EXPENDITURE
ON FOODGRAINS PER 30 DAYS

serial num- ber	fractile group (percent)	cumulative percentage of expenditure			cumulative percentage of expenditure on foodgrains		
		s.s.1	s.s.2	combined	s.s.1	s.s.2	combined
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	1	0.25	0.32	0.19	0.37	0.38	0.32
2	2	0.65	0.38	0.49	0.86	0.62	0.74
3	3	0.88	0.68	0.80	0.22	1.08	1.21
4	4	1.25	1.11	1.17	1.75	1.91	1.85
5	0—5	1.69	1.39	1.50	2.59	2.43	2.44
6	5—10	3.88	3.25	3.54	5.80	5.33	5.54
7	10—15	5.91	5.37	5.62	8.68	8.48	8.55
8	15—20	8.32	8.03	8.13	11.67	12.24	11.85
9	20—25	10.81	11.06	10.97	15.31	16.51	16.00
10	25—30	14.36	14.33	14.34	19.79	20.80	20.34
11	30—35	17.39	17.95	17.71	23.96	25.73	24.92
12	35—40	21.29	21.95	21.48	29.22	31.15	29.92
13	40—45	24.80	25.58	24.97	33.42	36.16	34.45
14	45—50	29.36	29.62	29.38	39.51	41.07	40.12
15	50—55	33.90	34.23	34.03	44.36	46.66	45.45
16	55—60	39.30	39.14	39.30	51.26	52.01	51.73
17	60—65	44.29	43.48	44.04	56.50	56.40	56.69
18	65—70	51.02	48.75	49.97	62.22	61.96	62.17
19	70—75	56.55	55.30	55.76	67.67	68.51	67.92
20	75—80	62.69	62.40	62.45	73.32	75.32	74.22
21	80—85	70.40	67.71	69.12	80.01	80.58	80.33
22	85—90	79.09	75.85	76.74	87.00	87.52	86.66
23	90—95	88.53	86.29	87.31	94.32	93.82	94.09
24	96	90.55	87.79	88.87	95.37	94.81	95.04
25	97	92.39	90.18	91.44	96.55	95.82	96.39
26	98	94.71	93.98	94.06	97.60	97.60	97.37
27	99	98.44	96.34	97.16	99.41	98.71	98.90
28	100	100.00	100.00	100.00	100.00	100.00	100.00

*The number of sample villages was 772 (including 4 uninhabited mauzas) and the number of sample households was 768 in each sub-sample.

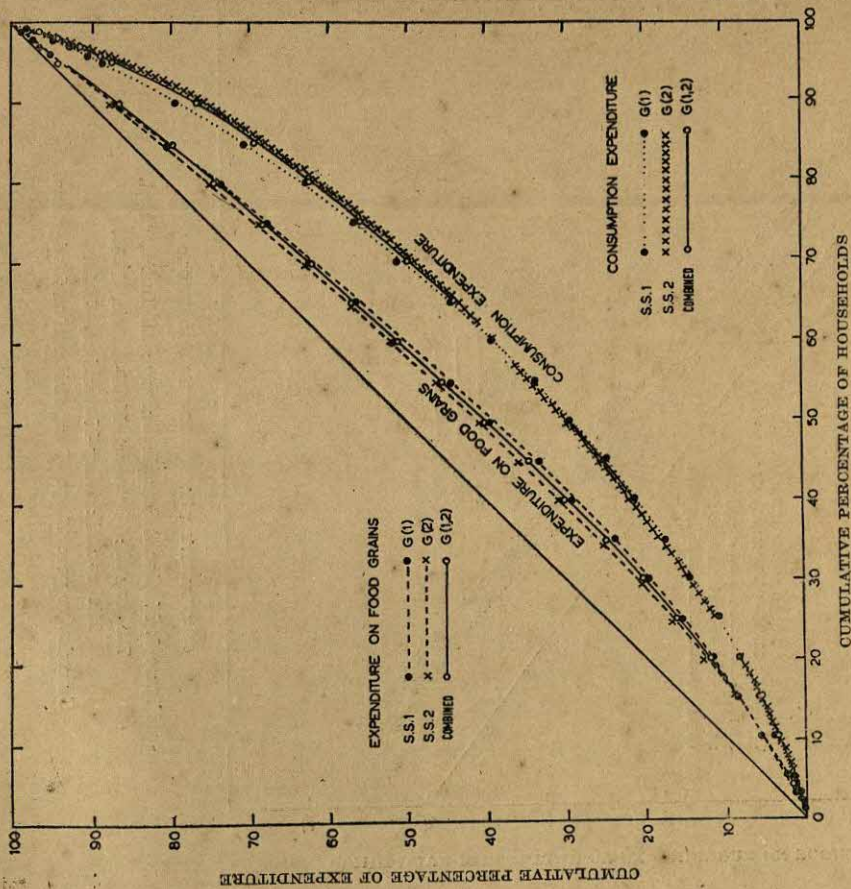


Figure 3. National Sample Survey: all India, rural.

7th round: October, 1953-March 1954.

Cumulative percentage of total consumer expenditure and expenditure on foodgrains per 30 days.

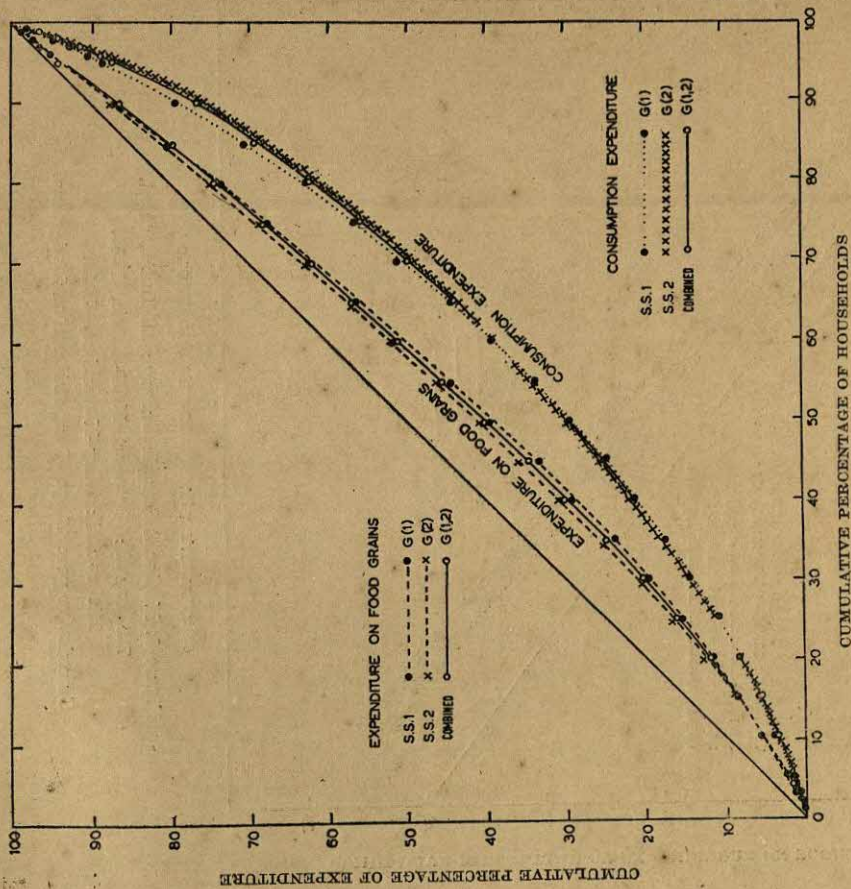


Figure 4. National Sample Survey: all India, rural.

9th round: May-November, 1955.

Cumulative percentage of total consumer expenditure and expenditure on foodgrains per 30 days.

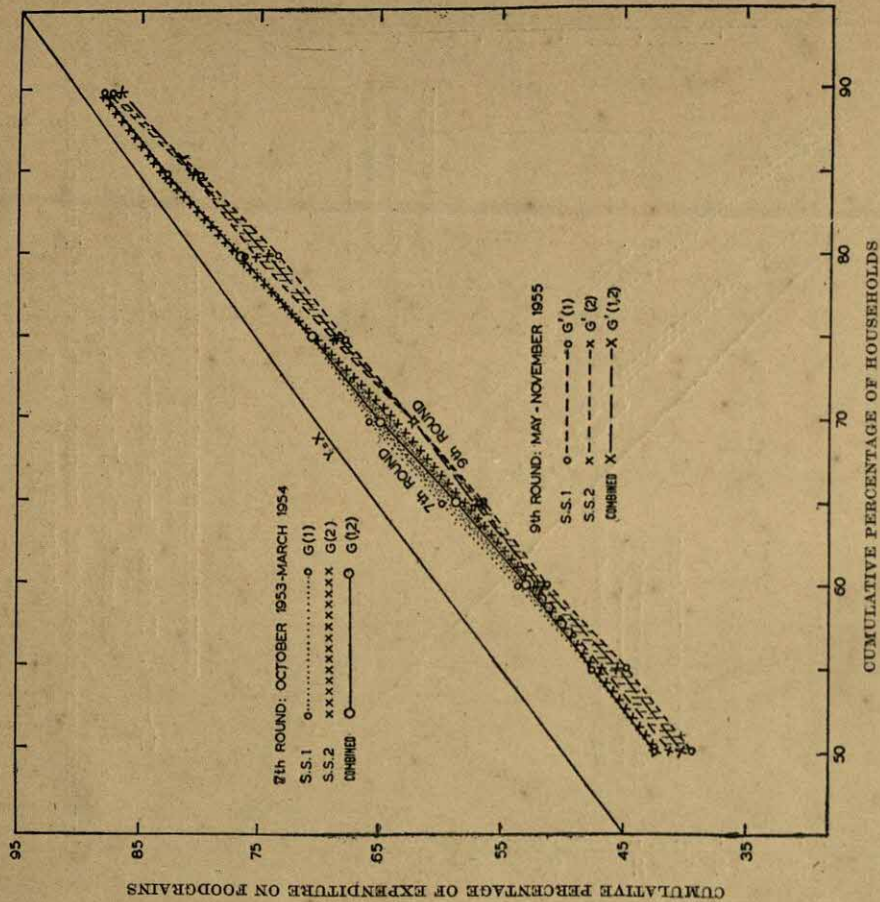


Figure 5. National Sample Survey: all India, rural.

7th round: October, 1953-March, 1954.

9th round: May-November, 1955.

Cumulative percentage of total expenditure and expenditure on foodgrains per 30 days.

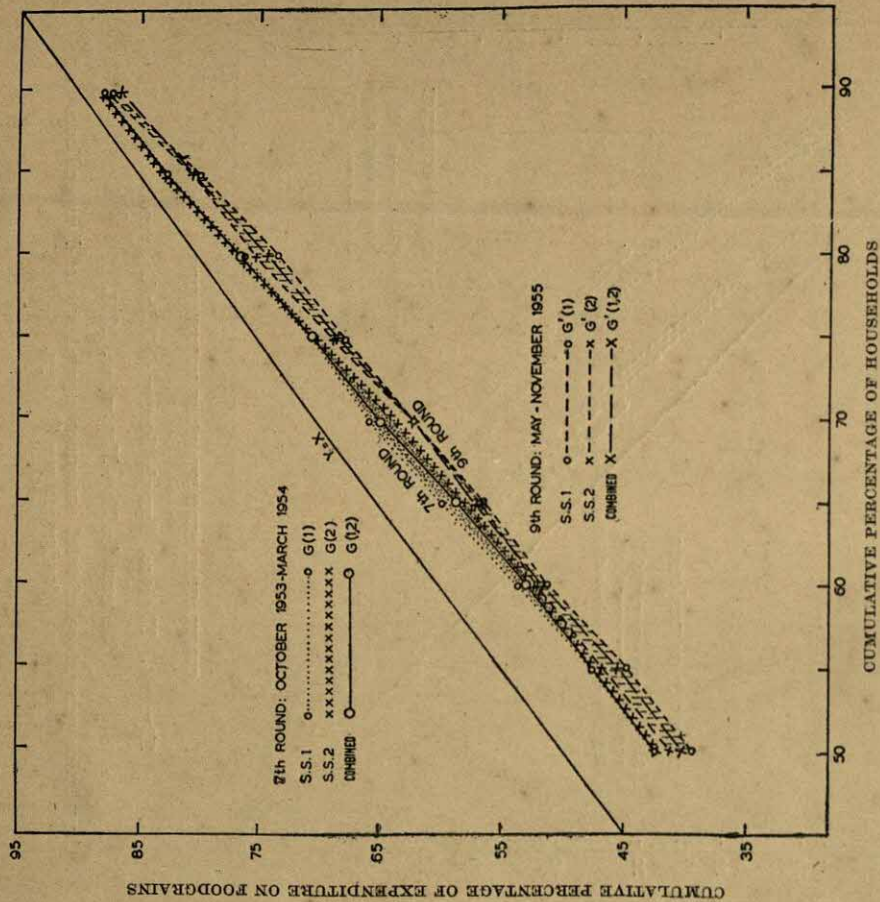


Figure 6. National Sample Survey: all India, rural.

7th round: October, 1953-March, 1954.

9th round: May-November, 1955.

Cumulative percentage of expenditure on foodgrains.

A METHOD OF FRACTILE GRAPHICAL ANALYSIS

3. REGIONAL DIFFERENCES

3.1. Consider a second example. In the 8th round (July, 1954—April, 1955) of the National Sample Survey information was collected for the whole of rural India (including Jammu and Kashmir) on holdings of land by households. Data for two States, namely, West Bengal and Andhra, may be used to illustrate the graphical method for comparisons between different geographical regions in the same period of time. There were twelve interpenetrating samples. Out of this material, consider only two sub-samples of 18 villages each from only two states, West Bengal and Andhra, with a total number of villages of 38,590 and 18,912, respectively. The number of sample households in West Bengal was 569 in the first sub-sample and 406 in the second sub-sample with a combined sample of 975 households from an estimated number of 5,413,000; the sampling fraction was about one in 5,600. In Andhra, the number of sample households was 343 and 360 in the two sub-samples, respectively, giving a combined sample of 703 out of an estimated number of 4,066,000 households; the over-all sampling fraction in this case was about one in 5,800. The variate selected for the present example is the land owned by each household in the rural area. There are, of course, many households owning no land, for example, households of landless labourers, artisans, professional people, etc., but they are included in the present study. The sample households were ranked by the size of their landholdings; and using appropriate probability weights, the number of households in the population and the area owned by each were estimated in the usual way. The estimated number of households was then divided into fractile groups and the land owned in each fractile group was also estimated. From this it is possible to calculate the cumulative percentage of owned area and also the average size of household ownership landholding in each fractile group. The data for West Bengal are shown in Table VII in which columns (2), (3), and (4) give the limiting value at the upper end of each fractile group of the size of ownership holdings in acres for the first sub-sample, the second sub-sample, and the combined sample, respectively. The next three columns—(5), (6), and (7)—give the cumulative percentage of land owned below and inclusive of each fractile group; and columns (8), (9), and (10) give the average size of a household ownership holding in each fractile group. Similar data are given for Andhra State in the corresponding columns of Table VIII.

3.2. Consider Figure 7.1 in which the x -axis represents the percentages of households (on the basis of ranking by amount of land owned) and the y -axis represents the average size of the land owned by households for each fractile group. The values for each of the five 20 per cent fractile groups have been plotted for both the sub-samples and the combined sample for both West Bengal and Andhra. The shaded area in each case gives the associated error. It can easily be seen that up to the 80th percentile there is overlapping of the error areas, showing that, on the basis of the available samples, the average size of holdings in the two States cannot be considered different. The average values for the top 20 per cent group are very clearly separated; and the average size of owned land for the top 20 per cent of households, as a whole, can be considered to be definitely higher in Andhra.

TABLE VII. NATIONAL SAMPLE SURVEY OF INDIA : RURAL
8TH ROUND : JULY, 1954-APRIL, 1955*
HOUSEHOLD OWNERSHIP HOLDINGS
State : West Bengal, Central Sample

fractile group	upper limit of size of ownership holding (acre)			cumulative percentage of owned area below			average size of household ownership holding (acres)		
	s.s.1	s.s.2	combined	s.s.1	s.s.2	combined	s.s.1	s.s.2	combined
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0-15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15-20	0.00	0.04	0.00	0.00	0.04	0.00	0.00	0.02	0.00
20-25	0.00	0.06	0.00	0.00	0.13	0.00	0.00	0.05	0.00
25-30	0.00	0.17	0.04	0.00	0.30	0.06	0.00	0.09	0.03
30-35	0.04	0.37	0.08	0.03	0.74	0.19	0.01	0.25	0.06
35-40	0.08	0.60	0.17	0.17	1.62	0.46	0.06	0.51	0.12
40-45	0.15	0.89	0.36	0.43	2.95	1.03	0.11	0.76	0.27
45-50	0.32	1.19	0.66	0.94	4.73	2.13	0.21	1.02	0.51
50-55	0.52	1.52	1.00	1.91	7.04	3.94	0.39	1.32	0.84
55-60	0.95	2.00	1.36	3.70	10.12	6.47	0.72	1.76	1.18
60-65	1.33	2.40	1.86	6.49	13.92	9.88	1.13	2.18	1.59
65-70	1.84	2.85	2.30	10.32	18.49	14.33	1.55	2.61	2.07
70-75	2.29	3.33	2.92	15.37	23.85	19.94	2.04	3.06	2.60
75-80	3.22	4.19	3.78	22.05	30.26	26.19	2.71	3.67	3.32
80-85	4.31	5.79	4.79	31.23	38.65	36.30	3.71	4.79	4.28
85-90	5.99	7.73	7.08	43.46	50.31	49.09	4.94	6.66	5.94
90-95	8.60	12.46	10.70	60.87	67.41	66.88	7.06	9.81	8.29
95-100	84.66	49.28	84.66	100.00	100.00	100.00	15.83	18.60	18.72
96	9.19	13.53	12.75	65.30	71.97	70.92	8.88	12.96	11.82
97	10.27	15.15	14.64	70.04	76.98	75.73	9.66	14.22	13.74
98	14.91	16.64	16.64	76.29	82.54	81.20	12.71	15.80	15.64
99	20.05	21.90	21.82	84.73	88.94	87.99	16.90	18.63	18.76
100	84.66	49.28	84.66	100.00	100.00	100.00	31.07	31.41	32.84
0-20	0.00	0.04	0.00	0.00	0.04	0.00	0.00	0.01	0.00
20-40	0.08	0.60	0.17	0.17	1.62	0.46	0.02	0.23	0.05
40-60	0.95	2.00	1.36	3.70	10.12	6.47	0.36	1.21	0.70
60-80	3.22	4.19	3.78	22.05	30.26	26.19	1.86	2.88	2.39
80-100	84.66	49.28	84.66	100.00	100.00	100.00	7.89	9.97	8.47
0-50	0.32	1.19	0.66	0.94	4.73	2.13	0.04	0.27	0.10
50-100	84.66	49.28	84.66	100.00	100.00	100.00	4.01	5.45	4.55
0-100	84.66	49.28	84.66	100.00	100.00	100.00	2.02	2.86	2.32

	sub-sample 1	sub-sample 2	combined
* number of sample villages	18	18	36
number of sample households	569	406	975

A METHOD OF FRACTILE GRAPHICAL ANALYSIS

TABLE VIII. NATIONAL SAMPLE SURVEY OF INDIA : RURAL
8TH ROUND : JULY, 1954—APRIL, 1955*
HOUSEHOLD OWNERSHIP HOLDINGS
State : Andhra, Central Sample

fractile group	upper limit of size of ownership holding (acre)			cumulative percentage of owned area below			average size of household ownership holding (acres)		
	s.s.1	s.s.2	combined	s.s.1	s.s.2	combined	s.s.1	s.s.2	combined
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0—15	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00
15—20	0.01	0.01	0.01	0.00	0.02	0.01	0.00	0.01	0.01
20—25	0.01	0.01	0.01	0.01	0.03	0.02	0.01	0.01	0.01
25—30	0.02	0.02	0.02	0.03	0.05	0.04	0.02	0.02	0.02
30—35	0.04	0.05	0.04	0.07	0.09	0.08	0.03	0.03	0.03
35—40	0.26	0.09	0.10	0.15	0.15	0.15	0.07	0.05	0.06
40—45	0.52	0.42	0.51	0.73	0.41	0.59	0.45	0.23	0.35
45—50	0.95	0.61	0.78	1.57	1.00	1.31	0.65	0.51	0.59
50—55	1.52	1.01	1.02	2.95	2.02	2.48	1.07	0.89	0.96
55—60	2.10	1.40	1.68	5.27	3.29	4.27	1.81	1.11	1.37
60—65	3.02	1.80	2.32	8.49	5.12	6.55	2.50	1.58	1.95
65—70	3.87	2.50	3.02	12.74	7.53	9.85	3.31	2.10	2.70
70—75	4.74	3.01	4.02	18.21	10.76	14.06	4.26	2.80	3.42
75—80	5.77	4.61	5.39	24.95	15.02	19.86	5.25	3.69	4.74
80—85	8.09	6.34	7.05	33.66	21.50	27.33	6.78	5.64	6.11
85—90	10.72	9.41	9.86	45.39	30.54	37.51	9.14	7.83	8.32
90—95	15.34	15.42	15.34	61.35	43.96	52.36	12.42	11.69	12.07
95—100	324.52	345.52	345.52	100.00	100.00	100.00	30.10	48.55	38.93
96	17.08	18.69	17.31	65.62	47.87	56.30	16.46	16.62	16.34
97	19.03	25.82	21.04	70.45	52.90	61.00	18.16	21.88	19.04
98	22.98	35.64	28.18	75.65	59.94	66.98	20.96	30.65	24.20
99	31.56	50.17	42.57	82.52	70.13	75.15	27.13	44.39	33.87
100	324.52	345.52	345.52	100.00	100.00	100.00	67.42	130.03	100.54
0—20	0.01	0.01	0.01	0.00	0.02	0.01	0.00	0.00	0.00
20—40	0.26	0.09	0.10	0.15	0.15	0.15	0.03	0.03	0.03
40—60	2.10	1.40	1.68	5.27	3.29	4.27	0.99	0.68	0.82
60—80	5.77	4.61	5.39	24.95	15.02	19.83	3.83	2.54	3.20
80—100	324.52	345.52	345.52	100.00	100.00	100.00	14.61	18.45	16.35
0—50	0.95	0.61	0.78	1.57	1.00	1.31	0.12	0.09	0.11
50—100	324.52	345.52	345.52	100.00	100.00	100.00	7.66	8.60	8.05
0—100	324.52	345.52	345.52	100.00	100.00	100.00	3.89	4.34	4.08

	sub-sample 1	sub-sample 2	combined
* number of sample villages	18	18	36
number of sample households	343	360	703

3.3. It is possible to make a more detailed examination of the range above 65 per cent of households. The average size of land owned in each 5 per cent fractile group between the 65th and 100th percentiles is plotted in Figure 7.2 for both West Bengal and Andhra; and the associated error areas are also shown in the usual way. In comparison with the associated error, the separation is significant for the group falling between the 90th and 95th percentiles, and is on the verge of significance for the fractile group between the 80th and 85th percentiles.

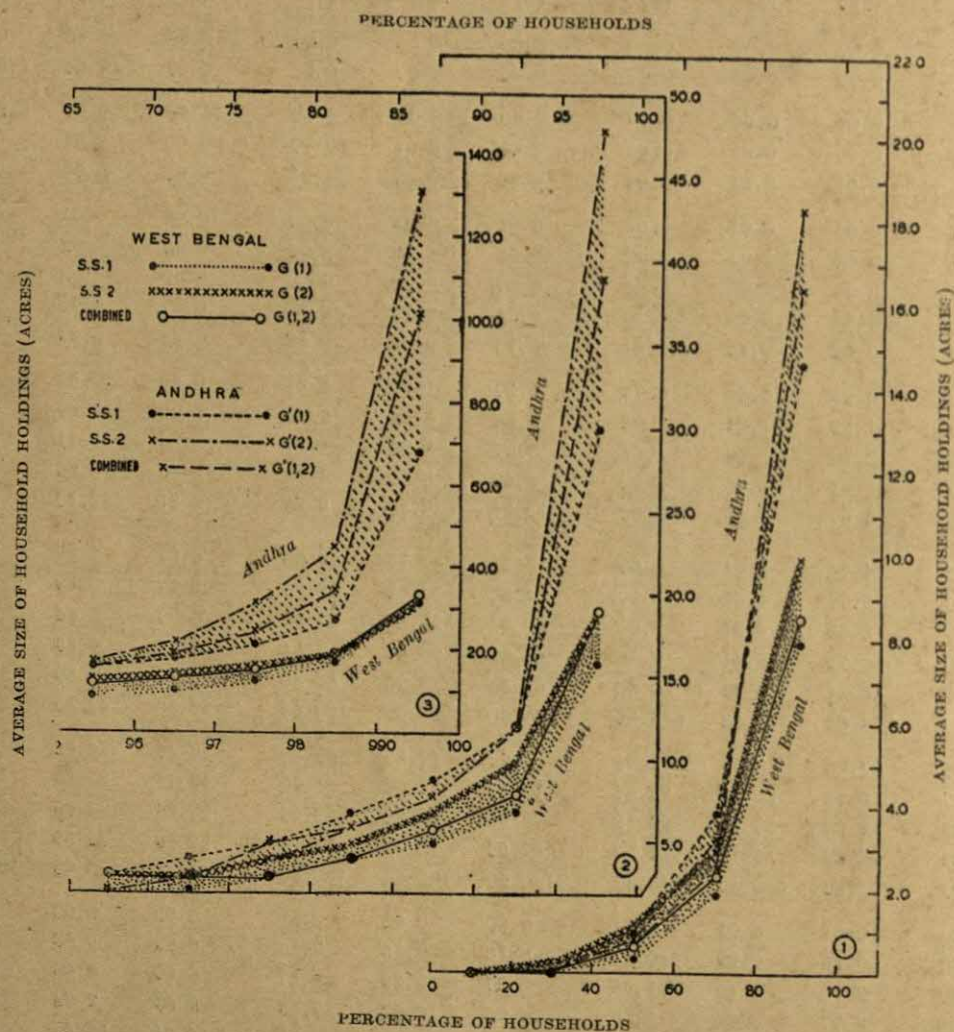


Figure 7.1-7.2-7.3. National Sample Survey of India, rural.

8th round: July, 1954-April, 1955.

Household ownership holding.

3.4. It is possible to go a step further and plot the average size of holdings for 1 per cent fractile groups between 90 and 100 per cent. This is shown in Figure 7.3. There is clear separation, even for 1 per cent groups, beyond 90 per cent of the households, but the separation becomes significant only at the level of the top one

A METHOD OF FRACTILE GRAPHICAL ANALYSIS

per cent of households. In considering these results, the extremely small size of the sample, only 18 villages in each sub-sample, must be kept in mind. The flexibility of the present method can, however, be easily appreciated.

4. COMPARISON WITH FREQUENCY DISTRIBUTIONS

4.1. It is worthwhile to contrast the present approach with that of the frequency distribution of the usual type in which the class ranges would be fixed, say, in terms of the money value of the per capita expenditure (such as Rs. 5, 10, 15, 20, etc.). In economic data, the money value of the expenditure would, in general, change with changes in price. The population of rural households as a whole in the 7th round may, of course, be compared in a meaningful way with the population of rural households as a whole in the 9th round. A fixed range frequency class would, however, represent different fractile groups in the two rounds and would not, therefore, be comparable in any important sense. The use of the *same* fractile groups would avoid conceptual difficulties involved in making comparisons over time or space or for two populations differing in any way. In the fractile approach, the bottom 10 per cent or the top 5 per cent, etc., of households in one round of the survey can be considered to be the counterpart of the bottom 10 per cent or the top 5 per cent, etc., of households in another round of the survey. This would also be true for comparisons between two States or two geographical regions for the same round. These fractile groups, in other words, may be treated as so many economic "strata" of the whole population and comparisons over time or space of the same stratum would be meaningful for many purposes.

4.2. The contrast can be expressed in a slightly different way. In using fixed ranges (with varying frequencies) the main interest lies in the pattern of the frequency distribution as a whole. In using equi-frequency fractile groups (with varying class intervals) it is possible to use each fractile group itself as a stratum or unit of comparison. The effect of price changes may, for example, be different at different levels of consumption expenditure, that is, in different fractile groups, and the picture for the population as a whole may become blurred. In the fractile method it is possible to study the effect for each fractile group separately, that is, to break up the whole spectrum of the range of expenditure into smaller and more homogeneous groups.

4.3. The fractile graphical approach offers an extremely rapid and practical method of analysis of statistical data of all types on a large scale. It is being used extensively now in the National Sample Survey of India. It would seem desirable to explore the possibilities of its applications and its usefulness in other fields.

In conclusion I should like to express my thanks to my colleagues in the Indian Statistical Institute who helped in the preparation of this paper.

Appendix

A METHOD OF FRACTILE GRAPHICAL ANALYSIS

The theoretical basis of the fractile graphical method is briefly explained in this appendix.

1. THEORETICAL BASIS

1.1. The *error* associated with the combined fractile graph $G(1, 2)$ is defined as the area lying between the two component sub-sample fractile graphs $G(1)$ and $G(2)$, and can be measured on the chart. This may be called e for the first population and e' for the second population.

1.2. For a non-graphical (that is, purely numerical) method of analysis, it is possible to define the error in either of the two usual forms, namely, (a) the sum of the differences (neglecting the sign of the difference) between the values of y for the two sub-sample graphs $G(1)$ and $G(2)$ for all fractile groups, or (b) the sums of the squares of these differences. All three were studied by model sampling experiments; and it was found that all three had similar distributions. The graphical definition has been selected, however, because of its simplicity. Once the accumulated totals are found (as explained in paragraphs 1.3—1.9 of the text) all the graphs can be easily drawn, and can also be interpreted directly by visual examination. In fact, junior computers can quickly learn this method of analysis, which makes it possible to use it on a very large scale. Secondly, the graphical method shows to what extent these two graphs cross and recross each other, and can sometimes reveal whether systematic non-sampling errors are present.¹

1.3. The combined fractile graph $G(1, 2)$ will usually lie partly within and partly outside the error area e ; and may occasionally lie entirely outside the error area e .

1.4. The *separation* has been defined as the area lying between the two combined graphs $G(1, 2)$ and $G'(1, 2)$.

¹ The presence of such non-sampling errors, of course, makes it impossible to give a rigorous mathematical theory, but this difficulty cannot be avoided as it is inherent in the method of a sample survey itself.

A METHOD OF FRACTILE GRAPHICAL ANALYSIS

2. RESULTS

2.1. A number of plausible results can now be stated.² Let N be the total number of sample units, g the number of fractile groups, and n the number of sample units in each fractile group, so that $N = gn$.

2.2. The combined graph $G(1, 2)$ would tend stochastically to lie entirely within the error area e as n increases when g is kept constant.

2.3. The error area e would tend to decrease inversely as $1/\sqrt{n}$ when g is kept constant, or as $\sqrt{(n_1 + n_2)/n_1 n_2}$ if the sizes of the two sub-samples are respectively n_1 and n_2 in each fractile group.

2.4. The error area e would tend to increase proportionately to g , the number of groups, when n remains the same (or when n_1 and n_2 the number of sample units for the two sub-samples in each fractile group, are kept constant).

2.5. Since $N = gn$, it follows that the error area e would tend to vary as $g^{3/2}$ when N is kept constant. If g is changed, and the new value of g is k times its original value, the relative changes in the error area would be approximately proportional to $k^{3/2}$. This is a most useful property as it enables one to use different values of g in testing the significance of the separation.

2.6. It is plausible to assume that the error (say, E) associated with the "separation" (say, S) would be given by $E^2 = (e)^2 + (e')^2$; and that (S^2/E^2) would tend to be distributed proportionally to χ^2 with g degrees of freedom.

2.7. When x and y are both random variates and are also statistically independent, the number of intersections of the two sub-sample graphs $G(1)$ and $G(2)$ would tend to be distributed like "runs" of heads and tails in g throws of an unbiased coin.

2.8. The above results would remain true for any set of linear and nonlinear transformations of the values of x and y in all the sub-samples.

3. SOME SUGGESTIONS FOR FURTHER WORK

3.1. The above results are, of course, only approximate, and the exact results would depend on the statistical distribution of both x and y and on the relationship between the two variates. For example, if n and g (and necessarily N , the size of the sub-sample) are kept the same, then the error area e for sampling from a population in which there is high correlation between y and x would tend to be smaller than the error area for sampling from a population in which y and x are less closely associated.

² These were first given in lectures at the Indian Statistical Institute in Calcutta in April, 1958 and at Berkeley, Chicago, and East Lansing in the United States in May and June, 1958. A preliminary note with the results of some model sampling experiments was published in the *Transactions of the Bose Institute*, Vol. XXII, 1958, pp. 223-230: "A Method of Fractile Graphical Analysis with Some Surmises of Results." Some further observations were made in lectures in Tokyo and Kiyushu, Japan in November and December, 1958.

3.2. Many model sampling experiments have been carried out, and the results are generally in agreement with the above conjectures within the margin of errors of sampling. Some of the results of the model sampling experiments have been published; other results are in the course of publication. The model sampling results present interesting suggestions. For example, the error area e may vary proportionately to $(g - \frac{1}{2})$ rather than to g , when n is kept constant. The number to be subtracted from g may not be exactly half, but some such correction factor would probably be necessary.

3.3. Results of model sampling experiments can show the plausibility of the conjectures but can never prove them in a theoretical sense. Theoretical investigations are, therefore, proceeding and some asymptotic results have been obtained which are broadly in agreement with the conjectures given above. A rigorous theory would open up many possibilities

ON SOME PROPERTIES OF ERROR AREA IN THE FRACTILE GRAPH METHOD

By K. TAKEUCHI

Faculty of Economics, Tokyo

SUMMARY: Some of the "surmises" given in Professor Mahalanobis' lecture delivered on 22 November 1958 at Tokyo University are investigated and proved.

This paper consists of three parts. In section 1, the expectation and variance of the "error area" $a(1, 2)$ are computed, under the condition that y_i ($i = 1, 2, \dots, g$) be normally distributed. In section 2, under rather general conditions, it is proved that y_i 's are asymptotically normally distributed. In section 3, special cases where x and y are bivariate normally distributed are investigated, and the computational results are compared with the experiments.

1. EXPECTATION AND VARIANCE OF THE ERROR AREA

The definitions and notations used in this paper will be same as in Mahalanobis (1958) which is much the same as in Mahalanobis (1960).

We first consider the error area between i -th and $i+1$ -th coordinates. We shall call it the i -th unit area and denote it by A_i . We assume that the distance between two coordinates is equal to 1.

We put

$$\begin{aligned} u &= \bar{y}_i - \bar{y}'_i \\ v &= \bar{y}_{i+1} - \bar{y}'_{i+1} \end{aligned}$$

and assume that u and v are normally distributed with means 0, common variance σ^2 , and correlation ρ .

Then,

$$\begin{aligned} A_i &= \frac{|u| + |v|}{2} \quad uv > 0 \\ &= \frac{u^2 + v^2}{2\{|u| + |v|\}} \quad uv < 0. \end{aligned}$$

Thus,

$$\begin{aligned} E(A_i) &= \int_0^\infty \int_0^\infty \frac{|u| + |v|}{2} p(u, v) du dv + \int_{-\infty}^0 \int_{-\infty}^0 \frac{|u| + |v|}{2} p(u, v) du dv + \\ &\quad + \int_{-\infty}^0 \int_0^\infty \frac{u^2 + v^2}{2\{|u| + |v|\}} p(u, v) du dv + \int_0^\infty \int_{-\infty}^0 \frac{u^2 + v^2}{2\{|u| + |v|\}} p(u, v) du dv \\ &= \int_0^\infty \int_0^\infty (u+v) p(u, v) du dv + \int_{-\infty}^0 \int_0^\infty \frac{u^2 + v^2}{u-v} p(u, v) du dv \end{aligned}$$

where $p(u, v)$ denotes the density of bivariate normal distribution.

Putting
$$r \cos \theta = \frac{u+v}{\sqrt{2(1+\rho)}}, \quad r \sin \theta = \frac{u-v}{\sqrt{2(1-\rho)}},$$

$$\alpha = \tan^{-1} \sqrt{\frac{1+\rho}{1-\rho}},$$

we obtain

$$\begin{aligned} \int_0^\infty \int_0^\infty (u+v) p(u, v) du dv &= \frac{1}{2\pi\sigma^2} \int_{-\alpha}^\alpha \int_0^\infty \sqrt{2(1+\rho)} \cos \theta \cdot r^2 e^{-r^2/2\sigma^2} dr d\theta \\ &= \frac{\sigma\sqrt{1+\rho}}{\sqrt{\pi}} \sin \alpha = \frac{\sigma(1+\rho)}{\sqrt{2\pi}} \end{aligned}$$

$$\begin{aligned} \int_{-\infty}^0 \int_0^\infty \frac{u^2+v^2}{u-v} p(u, v) du dv &= \frac{1}{2\pi\sigma^2} \int_\alpha^{\pi-\alpha} \int_0^\infty \frac{(1+\rho) \cos^2 \theta + (1-\rho) \sin^2 \theta}{\sqrt{2(1-\rho)} \sin \theta} r^2 e^{-r^2/2\sigma^2} dr d\theta \\ &= \frac{\sigma}{4\sqrt{\pi(1-\rho)}} \int_\alpha^{\pi-\alpha} \left(\frac{1+\rho}{\sin \theta} - 2\rho \sin \theta \right) d\theta \\ &= \frac{\sigma(1+\rho)}{2\sqrt{\pi(1-\rho)}} \log \frac{\sqrt{2} + \sqrt{1-\rho}}{\sqrt{1+\rho}} - \frac{\rho\sigma}{\sqrt{2\pi}} \\ E(A_i) &= \left\{ \frac{1+\rho}{2\sqrt{\pi(1-\rho)}} \log \frac{\sqrt{2} + \sqrt{1-\rho}}{\sqrt{1+\rho}} + \frac{1}{\sqrt{2\pi}} \right\} \sigma. \quad \dots (1.1) \end{aligned}$$

Therefore, $E(A_i) = \left(\frac{1}{2\sqrt{\pi}} \log (1+\sqrt{2}) + \frac{1}{\sqrt{2\pi}} \right) \sigma = 0.6475\sigma$ when $\rho = 0$.

$$E(A_i) = \sqrt{\frac{2}{\pi}} \sigma = 0.7979\sigma \quad \text{when } \rho = 1.$$

$$E(A_i) = \frac{1}{\sqrt{2\pi}} \sigma = 0.3980\sigma \quad \text{when } \rho = -1.$$

Now,
$$\begin{aligned} E(A_i^2) &= 2 \int_0^\infty \int_0^\infty \left(\frac{u+v}{2} \right)^2 p(u, v) du dv + 2 \int_{-\infty}^0 \int_0^\infty \left\{ \frac{u^2+v^2}{2(u-v)} \right\}^2 p(u, v) du dv \\ &= \frac{1+\rho}{2\pi\sigma^2} \int_{-\alpha}^\alpha \int_0^\infty \cos^2 \theta r^3 e^{-r^2/2\sigma^2} dr d\theta + \\ &\quad + \frac{1}{8\pi\sigma^2} \int_\alpha^{\pi-\alpha} \int_0^\infty \frac{\{(1+\rho) \cos^2 \theta + (1-\rho) \sin^2 \theta\}^2}{(1-\rho) \sin^2 \theta} r^3 e^{-r^2/2\sigma^2} dr d\theta \end{aligned}$$

$$\begin{aligned}
&= \frac{\sigma^2(1+\rho)}{\pi} \int_{-\alpha}^{\alpha} \cos^2 \theta \, d\theta + \frac{\sigma^2}{4\pi(1-\rho)} \int_{\alpha}^{\pi-\alpha} \left(\frac{(1+\rho)^2}{\sin^2 \theta} + 4\rho^2 \sin^2 \theta - 4\rho(1+\rho) \right) d\theta \\
&= \frac{\sigma^2(1+\rho)}{\pi} \left(\alpha + \frac{\sin 2\alpha}{2} \right) + \frac{\sigma^2}{4\pi(1-\rho)} \{ 2(1+\rho)^2 \cot \alpha + 2\rho^2(\pi - 2\alpha) + \\
&\quad + 2\rho^2 \sin 2\alpha - 4\rho(1+\rho)(\pi - 2\alpha) \} \\
&= \frac{\sigma^2(1+\rho)}{\pi} \left(\tan^{-1} \sqrt{\frac{1+\rho}{1-\rho}} + \frac{\sqrt{1-\rho^2}}{2} \right) + \\
&\quad + \frac{\sigma^2}{2\pi(1-\rho)} \left\{ (1+\rho+\rho^2)\sqrt{1-\rho^2} - \rho(2+\rho) \left(\pi - 2 \tan^{-1} \sqrt{\frac{1+\rho}{1-\rho}} \right) \right\} \quad \dots \quad (1.2)
\end{aligned}$$

$$V(A_i) = E(A_i^2) - \{E(A_i)\}^2$$

When $\rho = 0$, $V(A_i) = \left(\frac{1}{4} + \frac{1}{\pi} \right) \sigma^2 - \{E(A_i)\}^2 = 0.1491\sigma^2$.

Thus when we designate the error area by $a(1.2)$, we have,

$$E\{a(1.2)\} = \sum_{i=1}^{g-1} E(A_i).$$

But to compute the variance of $a(1.2)$, it is necessary to compute the covariance of A_i and A_{i+1} , which will be done later.

If we alter the definition of the area, and put $A'_i = \frac{|u| + |v|}{2}$ and write

$$a'(1.2) = \sum_{i=1}^{g-1} A'_i \text{ the distance area, then,}$$

$$E(A'_i) = \sqrt{\frac{2}{\pi}} \sigma = 0.7979\sigma$$

$$\begin{aligned}
E(A_i'^2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\frac{|u| + |v|}{2} \right)^2 p(u, v) \, du \, dv \\
&= 2 \int_{-\infty}^{\infty} \int_{-\infty}^{-\infty} \left(\frac{u+v}{2} \right)^2 p(u, v) \, du \, dv + 2 \int_{-\infty}^0 \int_0^{\infty} \left(\frac{u-v}{2} \right)^2 p(u, v) \, du \, dv \\
&= \frac{1+\rho}{2\pi\sigma^2} \int_{-\alpha}^{\alpha} \cos^2 \theta \, d\theta \int_0^{\infty} r^3 e^{-r^2/2\sigma^2} \, dr + \frac{1-\rho}{2\pi\sigma^2} \int_0^{\pi-\alpha} \sin^2 \theta \, d\theta \int_0^{\infty} r^3 e^{-r^2/2\sigma^2} \, dr \\
&= \frac{\sigma^2}{\pi} \left\{ (1+\rho) \left(\alpha + \frac{\sin^2 \alpha}{2} \right) + (1-\rho) \left(\frac{\pi+2\alpha}{2} + \frac{\sin 2\alpha}{2} \right) \right\} \\
&= \frac{\sigma^2}{\pi} \left\{ (1-\rho) \frac{\pi}{2} + 2\rho\alpha + \sin 2\alpha \right\} = \frac{\sigma^2}{\pi} \left\{ \frac{(1-\rho)}{2} \pi + 2\rho \tan^{-1} \sqrt{\frac{1+\rho}{1-\rho}} + \sqrt{1-\rho^2} \right\}.
\end{aligned}$$

In general,

$$E(A'_i) \geq E(A_i)$$

$$V(A'_i) \geq V(A_i)$$

But when $\rho = 0$,

$$E(A_i) = 0.6475\sigma$$

$$V(A_i) = 0.1491\sigma^2 \quad \dots (1.3)$$

$$E(A'_i) = 0.7979\sigma$$

$$V(A'_i) = 0.1817\sigma^2.$$

So that

$$\text{c.v. } (A_i) = 0.597 > \text{c.v. } (A'_i) = 0.535.$$

And when $\rho = 1$ always

$$A_i = A'_i.$$

If we assume that

$$\text{cov } (\bar{y}_i, \bar{y}_j) = 0, \quad i \neq j.$$

We have correlation

$$\rho(A'_i, A'_{i+j}) = 0, \quad j \geq 2$$

$$= \frac{1}{2}, \quad j = 1.$$

And if we assume

$$V(\bar{y}_i) = \frac{\sigma^2}{2} \quad (i = 1, 2, \dots).$$

We have,

$$\begin{aligned} V(a'(1.2)) &= V\left(\sum_{j=1}^{g-1} A'_j\right) \\ &= \Sigma V(A'_i) + 2\Sigma \text{cov } (A'_i, A'_j) \\ &= V(A')(2g-3) \end{aligned}$$

$$\text{c.v. } (a'(1.2)) \doteq \sqrt{\frac{2}{g}} \text{c.v. } (A'). \quad \dots (1.4)$$

Similarly when

$$V(\bar{y}_i) = \frac{\sigma^2}{2}, \text{cov } (\bar{y}_i, \bar{y}_j) = 0, \quad (i \neq j);$$

we have

$$\rho = 0 \quad \text{and} \quad \rho(A_i, A_{i+j}) = 0, \quad j \geq 2.$$

And

$$\begin{aligned} E(A_i A_{i+1}) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} A_i A_{i+1} p(u)p(v)p(w) du dv dw \\ &= 2 \int_0^{\infty} \int_0^{\infty} \int_0^{\infty} \frac{(u+v)(v+w)}{4} p(u)p(v)p(w) du dv dw \\ &\quad + 4 \int_{-\infty}^0 \int_0^{\infty} \int_0^{\infty} \frac{(u+v)(v^2+w^2)}{4(V-W)} p(u)p(v)p(w) du dv dw \\ &\quad + 2 \int_{-\infty}^0 \int_0^{\infty} \int_{-\infty}^0 \frac{(u^2+v^2)(v^2+w^2)}{4(v-w)} p(u)p(v)p(w) du dv dw \\ &= \frac{1}{2} I_1 + I_2 + \frac{1}{2} I_3 \text{ (say).} \end{aligned}$$

$$\begin{aligned}
 I_1 &= \int_0^\infty \int_0^\infty \int_0^\infty (v^2 + uv + vw + uv) p(u) p(u) p(v) p(w) du dv dw \\
 &= \frac{1}{4} \int_0^\infty v^2 p(v) dv + \frac{3}{2} \int_0^\infty \int_0^\infty uv p(u) p(v) du dv \\
 &= \frac{\sigma^2}{8} \left(1 + \frac{6}{\pi} \right). \quad \dots (1.5)
 \end{aligned}$$

$$\begin{aligned}
 I_2 &= \int_0^\infty \int_0^\infty \left(\frac{v^2 + w^2}{v + w} \right) p(u) p(w) dv dw \int_0^\infty (u + (v)) p(u) du \\
 &= \frac{1}{2} \int_0^\infty \int_0^\infty \left(\frac{v^2 + w^2}{v + w} \right) \left(\sqrt{\frac{2}{\pi}} \sigma + v \right) p(v) p(w) dv dw.
 \end{aligned}$$

By the transformation : $v = r \cos \theta$ $w = r \sin \theta$, we have

$$\begin{aligned}
 I_2 &= \frac{1}{2} \int_0^{\pi/2} \int_0^\infty \frac{1}{\cos \theta + \sin \theta} \left(\sqrt{\frac{2}{\pi}} \sigma + r \cos \theta \right) \frac{1}{2\pi\sigma^2} r^2 e^{-r^2/2\sigma^2} dr d\theta \\
 &= \frac{\sigma^2}{4\pi} \int_0^{\pi/2} \frac{d\theta}{\cos \theta + \sin \theta} + \frac{\sigma^2}{2\pi} \int_0^{\pi/2} \frac{\cos \theta d\theta}{\cos \theta + \sin \theta} \\
 &= \frac{\sigma^2}{2\pi} \left(\frac{1}{\sqrt{2}} \log (1 + \sqrt{2}) + \frac{\pi}{4} \right).
 \end{aligned}$$

I_3 is difficult to compute exactly, but

$$\begin{aligned}
 I_3 &= \int_0^\infty \int_0^\infty \int_0^\infty \frac{(u^2 + v^2)(v + w^2)}{(u + v)(v + w)} p(u) p(v) p(w) du dv dw \\
 &\leq \left\{ \int_0^\infty \int_0^\infty \int_0^\infty \left(\frac{v^2 + v^2}{u + v} \right)^2 p(u) p(v) p(w) du dv dw \right\}^{\frac{1}{2}} \times \\
 &\quad \times \left\{ \int_0^\infty \int_0^\infty \int_0^\infty \left(\frac{v^2 + w^2}{v + w} \right)^2 p(u) p(v) p(w) du dv dw \right\}^{\frac{1}{2}} \\
 &= \frac{1}{2} \int_0^\infty \int_0^\infty \left(\frac{v^2 + v^2}{u + v} \right)^2 p(u) p(v) du dv \\
 &= \frac{1}{2} \int_0^{\pi/2} \int_0^\infty \frac{1}{(\cos \theta + \sin \theta)^2} \frac{1}{2\pi\sigma^2} r^3 e^{-r^2/\sigma^2} dr d\theta = \frac{\sigma^2}{2\pi}.
 \end{aligned}$$

And by putting $v = v$, $u = v \tan \phi$, $w = v \tan \psi$,

$$\begin{aligned}
 I_3 &= \frac{1}{(2\pi\sigma^2)^{\frac{3}{2}}} \int_0^{\pi/2} \int_0^{\pi/2} \int_0^\infty \frac{\sec^2 \phi \sec^2 \psi}{(1 + \tan \phi)(1 + \tan \psi)} v^2 \sec^2 \phi \sec^2 \psi \times \\
 &\quad \times e^{-\frac{(1 + \tan^2 \phi + \tan^2 \psi) n^2}{2\sigma^2}} dv d\phi d\psi \\
 &= \frac{3\sigma^2}{4\pi} \int_0^{\pi/2} \int_0^{\pi/2} \frac{\sec^4 \phi \sec^4 \psi d\phi d\psi}{(1 + \tan \phi)(1 + \tan \psi)(1 + \tan^2 \phi + \tan^2 \psi)^{\frac{5}{2}}} \\
 &\geq \frac{3\sigma^2}{4\pi} \int_0^{\pi/2} \int_0^{\pi/2} \frac{d\phi d\psi}{(1 + \tan \phi)(1 + \tan \psi) \sec \phi \sec \psi} \\
 &= \frac{3\sigma^2}{4\pi} \left(\int_0^{\pi/2} \frac{\cos^2 \phi d\phi}{\cos \phi + \sin \phi} \right)^2 = \frac{3\sigma^2}{4\pi} \left(\frac{1}{2} \int_0^{\pi/2} \frac{d\phi}{\cos \phi + \sin \phi} \right)^2 \\
 &= \frac{3\sigma^2}{8\pi} (\log(1 + \sqrt{2}))^2. \tag{1.6}
 \end{aligned}$$

Thus we obtain,

$$\begin{aligned}
 &\left\{ \frac{3}{16} + \frac{3}{8\pi} + \frac{1}{2\sqrt{2}\pi} \log(1 + \sqrt{2}) + \frac{3}{16\pi} (\log(1 + \sqrt{2}))^2 \right\} \sigma^2 \\
 &\geq E(A_i A_{i+1}) \leq \left\{ \frac{3}{16} + \frac{5}{8\pi} + \frac{1}{2\sqrt{2}\pi} \log(1 + \sqrt{2}) \right\} \sigma^2,
 \end{aligned}$$

that is,

$$0.4524\sigma^2 \leq E(A_i A_{i+1}) \leq 0.4856\sigma^2.$$

Consequently,

$$\begin{aligned}
 \text{cov}(A_i, A_{i+1}) &= E(A_i A_{i+1}) - \{E(A_i)\}^2 \\
 &\leq 0.0663\sigma^2 \\
 &\geq 0.0331\sigma^2.
 \end{aligned}$$

$$0.222 \leq \rho(A_i, A_{i+1}) \leq 0.445.$$

This is not a good approximation, but is enough to see that $\rho(A_i, A_{i+1})$ is smaller than $\frac{1}{2}$ but is significantly larger than 0. To obtain a better approximation, we have to complete I_3 more exactly.

By putting

$$\rho(A_i, A_{i+1}) = \rho^*, \tag{1.7}$$

we have

$$V(a(1.2)) = V(A)\{(1 + 2\rho^*)g - (1 + 4\rho^*)\}.$$

For large g , we have approximately

$$\text{c.v.}(a(1.2)) = \sqrt{\frac{1 + 2\rho^*}{g}} \text{c.v.}(A) \leq \sqrt{\frac{2}{g}} \times 0.580 \geq \sqrt{\frac{2}{g}} \times 0.507. \tag{1.8}$$

ON SOME PROPERTIES OF ERROR AREA IN THE FRACTILE GRAPH METHOD

Remark added in proof: While Professor Mahalanobis was in Japan, I asked him if the distance area may be better than the error area as defined by him in view of the fact that $c.v.(A'_i)$ is smaller than $c.v.(A_i)$. The above shows that $c.v.(a'(1.2))$ need not be always smaller than $c.v.(a(1.2))$ because a small $\rho(A_i, A_{i+1})$ may compensate for a large $c.v.(A_i)$.

Remark: In the case when $V(\tilde{y}_i) \neq V(\tilde{y}_{i+1})$ by quite similar but somewhat lengthy calculations, we can have analogous results as above.

2. THE ASYMPTOTIC PROPERTY OF $\tilde{y}_i$

We assume the following model :

$$y = E(y|x) + \eta = f(x) + \eta$$

We put

$$z = f(x) \quad \text{so that} \quad y = z + \eta.$$

Assumption 1: η is distributed independently of x (so of z), and has the second order moment $\sigma_1^2 < \infty$.

Assumption 2: $f(x) = E(y|x)$ is a monotone increasing function of x .

Assumption 3: z has a continuous distribution with density $\phi(z)$ with a derivative and the second order moment.

Suppose that we have $n = n'g$ observations $(x, y_1), (x_2, y_2) \dots (x_n, y_n)$ and where $x_1 \leq x_2 \leq x_3 \leq \dots \leq x_n$ so that we obtain

$$(z_1, \eta_1), (z_2, \eta_2) \dots (z_n, \eta_n) \text{ where } z_1 \leq z_2 \leq \dots \leq z_n.$$

We put

$$\begin{aligned} \tilde{y}_i &= \frac{1}{n'} \sum_{j=1}^{n'} y_{n'(i-1)+j} \\ &= \frac{1}{n'} \sum_{j=1}^{n'} (z_{n'(i-1)+j} + \eta_{n'(i-1)+j}) \end{aligned}$$

and denote the same by

$$\bar{z}_i + \bar{\eta}_i.$$

By Assumption 1, $\bar{z}_i$ and $\bar{\eta}_i$ are independent and when n' is large, $\bar{\eta}_i$ is asymptotically normally distributed with mean 0, and variance $O\left(\frac{1}{n'}\right)$. To prove the asymptotic normality of $\bar{z}_i$, we utilize the following :

Lemma: Let $(x_1, \xi_1), (x_2, \xi_2) \dots$ be a sequence of pairs of random variables.

If, the conditional probability distribution of x_n given $\xi_n = \xi$ tends to a normal distribution, with mean $\mu(\xi)$, variance $\frac{1}{n} \sigma^2(\xi)$, for every value of ξ ; $\mu(\xi)$, $\sigma(\xi)$ are continuous functions of and have second order derivatives; and ξ_n tends to normality with mean $\bar{\xi}$, variance $\frac{1}{n} \sigma_{\xi}^2$. Then, x_n is asymptotically normally distributed, with mean μ , variance $\frac{1}{n} \sigma^2$, where

$$\begin{aligned} \mu &= \mu(\bar{\xi}) \\ \sigma^2 &= \sigma^2(\bar{\xi}) + (\mu'(\bar{\xi}))^2 \sigma_{\xi}^2. \end{aligned}$$

Proof: Let $\varphi_n(t) = E\{e^{itx_n}\}$.

By assumption,

$$\begin{aligned} E\{e^{itx_n} | \xi_n\} &= \exp \left[i\mu(\xi_n)t - \frac{1}{2n} \sigma^2(\xi_n)t^2 + o\left(\frac{1}{n}\right) \right] \\ &= \exp \left[i\mu(\bar{\xi})t + i\mu'(\bar{\xi})(\xi_n - \bar{\xi})t - \frac{1}{2n} \sigma^2(\bar{\xi})t^2 + \right. \\ &\quad \left. + O\left((\xi_n - \bar{\xi})^2\right) + O\left(\frac{1}{n}(\xi_n - \bar{\xi})\right) + o\left(\frac{1}{n}\right) \right] \\ &= \exp[i\mu(\bar{\xi})t - \frac{1}{2n} \sigma^2(\bar{\xi})t^2] \times \\ &\quad \times \exp \left[i\mu'(\bar{\xi})(\xi_n - \bar{\xi})t + O((\xi_n - \bar{\xi})^2) + O\left(\frac{1}{n}(\xi_n - \bar{\xi})\right) + o\left(\frac{1}{n}\right) \right] \\ \varphi_n(t) &= E\{E(e^{itx_n} | \xi_n)\} = \exp \left[i\mu(\bar{\xi})t - \frac{1}{2n} \sigma^2(\bar{\xi})t^2 \right] \times \\ &\quad \times E\{\exp \left[i\mu'(\bar{\xi})(\xi_n - \bar{\xi})t + O((\xi_n - \bar{\xi})^2) + O\left(\frac{1}{n}(\xi_n - \bar{\xi})\right) + o\left(\frac{1}{n}\right) \right]\} \\ &= \exp[i\mu(\bar{\xi})t - \frac{1}{2n} \sigma^2(\bar{\xi})t^2] \times \exp \left[-\frac{(\mu'(\bar{\xi}))^2}{2n} \sigma_{\xi}^2 t^2 + o\left(\frac{1}{n}\right) \right] \\ &= \exp \left[i\mu(\bar{\xi})t - \frac{1}{2n} \{\sigma^2(\bar{\xi}) + (\mu'(\bar{\xi}))^2 \sigma_{\xi}^2\} t^2 + o\left(\frac{1}{n}\right) \right]. \end{aligned}$$

which completes the proof.

Similarly, let $(x_1, \xi_1, \eta_1)(x_2, \xi_2, \eta_2) \dots$ be a sequence of random variables such that the conditional probability of x_n given $\xi_n = \xi$ and $\eta_n = \eta$ tends to a normal distribution with mean $\mu(\xi, \eta)$, variance $\frac{1}{n} \sigma^2(\xi, \eta)$ as $n \rightarrow \infty$, where μ and σ^2 is a continuous function having second order derivatives, and that (ξ_n, η_n) tends to a bivariate normal distribution $N(\bar{\xi}, \bar{\eta}; \sigma_{\xi}^2, \sigma_{\eta}^2, \sigma_{\xi\eta})$; then x_n is asymptotically normal with mean $\mu(\bar{\xi}, \bar{\eta})$ and variance

$$\frac{1}{n} \sigma^2 = \frac{1}{n} \{ \sigma^2(\bar{\xi}, \bar{\eta}) + (\mu_{\xi}(\bar{\xi}, \bar{\eta}))^2 \sigma_{\xi}^2 + (\mu_{\eta}(\bar{\xi}, \bar{\eta}))^2 \sigma_{\eta}^2 + 2\mu_{\xi}(\bar{\xi}, \bar{\eta})\mu_{\eta}(\bar{\xi}, \bar{\eta}) \sigma_{\xi\eta} \}$$

where

$$\mu_{\xi}(\bar{\xi}, \bar{\eta}) = \frac{\partial}{\partial \xi} \mu(\xi, \eta) \Big|_{\xi = \bar{\xi}, \eta = \bar{\eta}}$$

$$\mu_{\eta}(\bar{\xi}, \bar{\eta}) = \frac{\partial}{\partial \eta} \mu(\xi, \eta) \Big|_{\xi = \bar{\xi}, \eta = \bar{\eta}}$$

... (2.1)

ON SOME PROPERTIES OF ERROR AREA IN THE FRACTILE GRAPH METHOD

First we shall show the asymptotic normality of $\bar{z}_1$.

$$\bar{z}_1 = \frac{z_1 + z_2 + \dots + z_{n'}}{n'}$$

and given $z_{n'} = t$, $n'z_1$, is equal to the sum of constant t and $n'-1$ independent variables, which have the same distribution $P(z|z \leq t)$ with density

$$\frac{\phi(z)}{\int_{-\infty}^t \phi(z) dz}, \quad (z \leq t)$$

and since this probability distribution has the second-order moment, $\bar{z}_1$ is conditionally asymptotically normally distributed given $z_{n'} = t$, with mean and variance,

$$\mu(t) = \frac{\int_{-\infty}^t z\phi(z)dz}{\int_{-\infty}^t \phi(z)dz}, \quad \frac{1}{n'} \sigma^2(t) = \frac{1}{n'} \frac{\int_{-\infty}^t (z-\mu(t))^2 \phi(z)dz}{\int_{-\infty}^t \phi(z)dz}.$$

And since z_n can be regarded to be a $\frac{1}{g}$ fractile, it is asymptotically normal

($\mu = \zeta_1$, $\sigma = \frac{1}{\phi(\zeta_1)} \sqrt{\frac{g-1}{ng^2}}$) where ζ_1 is the $\frac{1}{g}$ fractile of the population of z , (Cramer (1946)).

Then, applying the above lemma, z_1 is asymptotically normal, with mean,

$$\mu = \mu(\zeta_1)$$

and since

$$\begin{aligned} \mu'(\zeta_1) = \mu'(t) |_{t=\zeta_1} &= \frac{t\phi(t) \int_{-\infty}^t \phi(z)dz - \phi(t) \int_{-\infty}^t z\phi(z)dz}{\left\{ \int_{-\infty}^t \phi(z)dz \right\}^2} \Big|_{t=\zeta_1} \\ &= g\phi(\zeta_1)(\zeta_1 - \mu(\zeta_1)) \end{aligned}$$

the variance is

$$\begin{aligned} \frac{1}{n'} \sigma_1^2 &= \frac{1}{n'} \sigma^2(\zeta_1) + \{g\phi(\zeta_1)(\zeta_1 - \mu(\zeta_1))\}^2 \frac{1}{\phi(\zeta_1)^2} \frac{g-1}{n'g^3} \\ &= \frac{1}{n'} \left\{ \sigma^2(\zeta_1) + \frac{g-1}{g} (\zeta_1 - \mu(\zeta_1))^2 \right\}. \end{aligned} \quad \dots (2.2)$$

As for $\bar{z}_i (i \geq 2)$, considering first conditionally given $z_{n'(i-1)+1}$ and $z_{n'i-1}$, then by utilizing the fact that $z_{n'(i-1)+1}$ and $z_{n'i-1}$ are asymptotically bivariate normal, we can show in quite a similar way as above, that $\bar{z}_i$'s are asymptotically normally distributed, with variance $\frac{1}{n'} \sigma_i^2$ and σ_i^2 is expressed as

$$\begin{aligned} \sigma_i^2 = & \sigma^2(\zeta_{i-1}, \zeta_i) + \frac{i(g-i)}{g} (\zeta_i - \mu(\zeta_{i-1}, \zeta_i))^2 + \\ & + \frac{(i-1)(g-i+1)}{g} (\mu(\zeta_{i-1}, \zeta_i) - \zeta_{i-1})^2 + \\ & + \frac{2(i-1)(g-i)}{g} (\zeta_i - \mu(\zeta_{i-1}, \zeta_i))(\mu(\zeta_{i-1}, \zeta_i) - \zeta_{i-1}) \quad \dots (2.3) \end{aligned}$$

where

$$\mu(\zeta_{i-1}, \zeta_i) = \frac{\int_{\zeta_{i-1}}^{\zeta_i} z \phi(z) dz}{\int_{\zeta_{i-1}}^{\zeta_i} \phi(z) dz} = g \frac{\int_{\zeta_{i-2}}^{\zeta_i} z \phi(z) dz}{\int_{\zeta_{i-2}}^{\zeta_i} \phi(z) dz}$$

$$\sigma^2(\zeta_{i-1}, \zeta_i) = \frac{\int_{\zeta_{i-1}}^{\zeta_i} (z - \mu(\zeta_{i-1}, \zeta_i))^2 \phi(z) dz}{\int_{\zeta_{i-1}}^{\zeta_i} \phi(z) dz} = g \frac{\int_{\zeta_{i-2}}^{\zeta_i} (z - \mu(\zeta_{i-1}, \zeta_i))^2 \phi(z) dz}{\int_{\zeta_{i-2}}^{\zeta_i} \phi(z) dz}.$$

Thus we obtain :

Theorem : Under Assumption 1-3, $\bar{y}_i$'s are asymptotically normally distributed, with variance $\frac{1}{n} (\sigma_i^2 + \sigma_\eta^2)$.

Proof : This is an immediate consequence of the asymptotic normality of $\bar{z}_i$ and $\bar{y}_i$.

Remark : The above assumption 2 is taken only to simplify the calculation; the monotonicity of $f(x)$ is neither necessary nor essential. Assumption 1 also may be weakened.

Remark added in proof : A generalization was made by Prof. Kitagawa in connection with these assumptions; he showed that the additive model of z is inessential for the results as given above.

ON SOME PROPERTIES OF ERROR AREA IN THE FRACTILE GRAPH METHOD

3. NUMERICAL EXAMPLES WHEN x AND y ARE BIVARIATE NORMAL

We consider the case where x and y are bivariate normal

$$N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho_{xy})$$

We put

$$z = \frac{\sigma_y}{\sigma_x} \rho_{xy} (x - \mu_x) + \mu_y$$

$$y = z + \eta$$

then z and η are independent, normal, with variances,

$$\sigma_z^2 = \rho_{xy}^2 \sigma_y^2 \quad \sigma_\eta^2 = (1 - \rho_{xy}^2) \sigma_y^2.$$

We shall take up the case of model sampling experiments in the Indian Statistical Institute (Mahalanobis, 1958).

In this case

$$\mu_x = \mu_y = 0, \quad \sigma_x = 1, \quad \sigma_y = \sqrt{2}, \quad \rho_{xy} = \frac{1}{\sqrt{2}}, \quad \sigma_z^2 = \sigma_\eta^2 = 1.$$

(1) *The case $g = 2$.* According to the notation of Section 2.

$$\zeta_1 = 0, \quad \phi(\zeta_1) = \frac{1}{\sqrt{2\pi}}, \quad \mu(\zeta_1) = -\sqrt{\frac{2}{\pi}}, \quad \sigma^2(\zeta) = 1 - \frac{2}{\pi}.$$

By (2.2), we obtain,

$$V(\bar{z}_1) = \frac{1}{n'} \left\{ \left(1 - \frac{2}{\pi} \right) + \frac{1}{\pi} \right\} = \frac{1}{n'} \left(1 - \frac{1}{\pi} \right)$$

$$V(\bar{y}_1) = V(\bar{z}_1) + V(\bar{\eta}_1) = \frac{1}{n'} \left(2 - \frac{1}{\pi} \right)$$

Obviously,

$$V(\bar{z}_1) = V(\bar{z}_2), \quad V(\bar{y}_1) = V(\bar{y}_2).$$

And since

$$\bar{z}_1 + \bar{z}_2 = \frac{1}{n'} \sum \bar{z}_i = 2\bar{z} \text{ (total sample mean)}$$

$$V(\bar{z}_1 + \bar{z}_2) = 2V(\bar{z}_1) + 2 \text{ cov}(\bar{z}_1, \bar{z}_2) = 4V(\bar{z}) = \frac{2}{n'}.$$

Consequently,

$$\text{cov}(\bar{z}_1, \bar{z}_2) = \frac{1}{n'} - \frac{1}{n'} \left(1 - \frac{1}{\pi} \right) = \frac{1}{n'\pi}$$

$$\text{cov}(\bar{y}_1, \bar{y}_2) = \text{cov}(\bar{z}_1, \bar{z}_2) = \frac{1}{n'\pi}$$

$$\rho(\bar{y}_1, \bar{y}_2) = \frac{\frac{1}{n'\pi}}{\frac{1}{n'} \left(2 - \frac{1}{\pi} \right)} = \frac{1}{2\pi - 1} = 0.1893.$$

Using the notation of Section 1,

$$\sigma_u^2 = \sigma_v^2 = V(\bar{y}_1 + \bar{y}'_1) = 2V(\bar{y}_1) = \frac{1}{n'} \left(4 - \frac{2}{\pi} \right)$$

$$\text{cov}(u, v) = \text{cov}(\bar{y}_1 - \bar{y}'_1, \bar{y}_2 - \bar{y}'_2) = 2 \text{cov}(\bar{y}_1, \bar{y}_2)$$

$$\rho_{uv} = \rho(\bar{y}_1, \bar{y}_2) = 0.1893$$

$$\sigma_u = \sigma_v = \frac{1}{n'} \times 1.834.$$

Then by (1.1) we have

$$E(a(1.2)) = E(A_1) = \frac{1}{\sqrt{n'}} \times 1.246.$$

And by (1.1) and (1.2)

$$\text{c.v.}(a(1.2)) = \text{c.v.}(A_1) = 0.627$$

which do not contradict with the results of the experiments in Mahalanobis (1958).

We state here some wellknown results (See Kendall and Stuart (1958)) which will be used in the remaining sections.

Let
$$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \quad \Phi(t) = \int_{-\infty}^t \phi(t) dt.$$

We have for $t > 0$

$$\frac{t}{t^2+1} < \frac{1-\Phi(t)}{\phi(t)} < \frac{1}{t} \quad \dots \quad (3.1)$$

$$\frac{1-\Phi(t)}{\phi(t)} = \frac{1}{t} - \frac{1}{t^3} + \frac{1.3}{t^5} - \dots (-1)^r \frac{1.3 \dots (2r-1)}{t^{2r+1}} + R_r(t)$$

where $|R_r(t)| < 1.3 \dots (2r-1) \cdot t^{-(2r+1)} \quad \dots \quad (3.2)$

$$\frac{1-\Phi(t)}{\phi(t)} < \sqrt{\frac{\pi}{2}} \quad \dots \quad (3.3)$$

since the expression on the left hand side is monotonically decreasing.

(2) *The case $g \geq 4$.* In this case also we can compute the variances and covariances of $\bar{z}_i$ and $\bar{y}_i$. Since the computations are rather lengthy, we shall give upper and lower bounds for these.

Let

$$2 \leq i \leq \frac{g-1}{2}$$

$$V(\bar{z}_i) < V(\bar{z}_{g(i-1)+1}) \doteq \frac{1}{g n'} \frac{1}{\phi(\zeta_{i-1})^2} \frac{(i-1)(g-i+1)}{g^2}$$

From (3.3) we see that

$$V(\bar{z}_i) < \frac{\pi\sigma^2(g-i+1)}{2n'g(i-1)} \quad \dots \quad (3.4)$$

Next

$$V\left(\frac{\bar{z}_i + \bar{z}_{g-i+1}}{2}\right) \geq V(\bar{z}) = \frac{\sigma^2}{gn'}$$

since $\frac{\bar{z}_i + \bar{z}_{g-i+1}}{2}$ is an unbiased estimator of μ_z of which $\bar{z}$ is the efficient estimate.

$$\text{Hence} \quad V(\bar{z}_i) \geq \frac{2\sigma^2}{gn'(1+\rho_i)} \quad \dots \quad (3.5)$$

where ρ_i is the correlation coefficient between $\bar{z}_i$ and $\bar{z}_{g-i+1}$ (3.4) and (3.5) furnish upper and lower bounds for $V(\bar{z}_i)$.

(3) *The case when g is large.* We first show that $n'V(\bar{z}_i)$ tends to zero uniformly in i as $g \rightarrow \infty$. Let $2 \leq i \leq g-1$.

From (2.3) we see that

$$\sigma_i^2 < (\zeta_i - \zeta_{i-1})^2 \left\{ 1 + \frac{i(g-1)}{g} + \frac{2(i-1)(g-i)}{g} + \frac{(i-1)(g-i+1)}{g} \right\}$$

For

$$2 \leq i \leq \frac{g-1}{2}, \text{ we have}$$

$$\frac{1}{g} = \int_{\zeta_{i-1}}^{\zeta_i} \phi(z) dz \geq (\zeta_i - \zeta_{i-1}) \phi(\zeta_{i-1})$$

Therefore,

$$\zeta_i - \zeta_{i-1} \leq \frac{1}{g\phi(\zeta_{i-1})} \leq \frac{\sigma}{|\zeta_{i-1}|(i-1)}$$

by an application of (3.1).

Thus

$$\sigma_i^2 < \frac{5\sigma^2}{|\zeta_{i-1}|^2} \quad \dots \quad (3.6)$$

If $i = g$ we have from (2.3)

$$\sigma^2 = g \int_{\zeta_g}^{\infty} z^2 \frac{e^{-z^2/2\sigma^2}}{\sqrt{2\pi}\sigma} dz - \mu_g^2 + \frac{g-1}{g} (\mu_g - \zeta_g)^2$$

where

$$\mu_g = \sigma \frac{g}{\sqrt{2\pi}} e^{-\zeta_g^2/2\sigma^2}$$

We also see that

$$g \int_{\zeta_g}^{\infty} z^2 \frac{e^{-z^2/2\sigma^2}}{\sqrt{2\pi}\sigma} dz = \mu_g \zeta_g + \sigma^2.$$

Using (3.2) with $r = 2$, we see that

$$\sigma_g^2 < \frac{\text{const}}{\xi^2} \quad \dots (3.7)$$

We further note that $\sigma_i^2 = \sigma_{g-i+1}^2$.

We easily see that for a fixed k and any $\epsilon > 0$ we can find a G such that for all $g > G$.

$$\sigma_i^2 < \epsilon \text{ if } 2 \leq i \leq \frac{k}{\sqrt{g}} \text{ or } g - \frac{k}{\sqrt{g}} \leq i \leq g-1 \text{ from (3.6)}$$

$$\sigma_i^2 < \epsilon \text{ if } \frac{k}{\sqrt{g}} < i < g - \frac{k}{\sqrt{g}} \text{ from (3.4)}$$

and $\sigma_i^2 < \epsilon$ if $i = 1$ or g from (3.7).

so that $\sigma_i^2 \rightarrow 0$ uniformly in i as $g \rightarrow \infty$.

From this we see that $n'V(\bar{y}_i) = n'V(\bar{z} + \bar{\eta}_i) \rightarrow \sigma_\eta^2$ uniformly in i and $\text{cor}(\sqrt{n'}V(\bar{y}_i), \sqrt{n'}V(\bar{y}_j)) \rightarrow 0$ uniformly in i and j , so that, from (1.3) we have

$$E(a(1, 2)) = 0.6475 \times \sigma \times \frac{g-1}{\sqrt{n'}} = \frac{0.916 \times (g-1)}{\sqrt{n'}} \quad \dots (3.8)$$

And, using (1.4) we have

$$\text{c.v.}(a(1, 2)) = \sqrt{\frac{2}{g}} \times 0.51. \quad \dots (3.9)$$

Both these relations are in fair agreement with the model sampling experiments.

I am much indebted to Professor Masuyama for encouragement and advices and to Mr. M. Shibuya, the discussion with whom was very suggestive and useful. I am also grateful to Miss T. Hayashi and Miss Matsui for preparing this manuscript.

REFERENCES

- CRAMÉR, H. (1946): *Mathematical Methods of Statistics*, Princeton University Press, 369.
 KENDALL, M. G. and STUART, A. (1958): *The Advanced Theory of Statistics*, Vol. I. Distribution Theory, Charles Griffin and Company, Ltd.
 MAHALANOBIS, P. C. (1958): Lectures in Japan: Fractile Graphical Analysis, Indian Statistical Institute.
 ——— (1960): A method of Fractile Graphical Analysis. *Econometrica*, 28, 2, 325-351, reprinted in *Sankhyā*, Series A, 23, 41-64.

Paper received: June, 1959.

Revised: March, 1960.

SOME LIMIT DISTRIBUTIONS CONNECTED WITH FRACTILE GRAPHICAL ANALYSIS

By J. SETHURAMAN

Indian Statistical Institute

SUMMARY. Fractile graphs were introduced by Professor P. C. Mahalanobis (1958), (1960)* for describing certain features of bivariate populations. The problem is to make useful comparisons between two bivariate populations wherein the values of one of the variables, suspected to be inflated or deflated by measurement, is omitted and its rankings alone are taken into account. Professor Mahalanobis (see Mahalanobis (1958)) introduced the error area as a measure of divergence between two fractile graphs and made some surmises concerning its behaviour. In this paper we introduce other measures of divergence and investigate into their limiting distributions in the light of his surmises (see Mahalanobis (1958)).

1. INTRODUCTION

1.1. We deal only with some theoretical aspects of Fractile Graphical Analysis. In our next note, we hope to deal with testing and other practical aspects of Fractile Graphical Analysis.

The summary of all our results in this paper is given in Section 3 after the definitions and notations are dealt with in Section 2.

2. DEFINITIONS AND NOTATIONS

2.1. Let $(y_1, x_1), \dots, (y_n, x_n)$ be a sample of n independent observations on a random variable (Y, X) . Let $n = m \cdot g$ where m and g are integers. Rearrange the sample thus

$$(y_{(1)}, x_{(1)}), \dots, (y_{(n)}, x_{(n)})$$

so that

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

Let

$$u_i = \frac{1}{m} \sum_{r=(i-1)m+1}^{im} x_{(r)}; \quad i = 1, \dots, g \quad \dots (2.1.1)$$

$$v_i = \frac{1}{m} \sum_{r=(i-1)m+1}^{im} y_{(r)}.$$

Thus from a sample we get a sequence $(v_1, \dots, v_g)$. Plotting these against the equidistant points $1, \dots, g$ and joining the points successively, we get a curve G known as the fractile graph of the samples.

2.2. In the problem we are to discuss we will deal with several samples. Each sample will bear an index. The statistics from a sample with index (i) are

* Fractile Graphical Analysis was first introduced in a lecture at the Indian Statistical Institute, Calcutta, in April 1958 and developed at Berkeley, Chicago, and at other places in USA in May and June 1958, and later on in Japan.

derived from those of the unindexed sample by affixing the index (i), for example, G^i denotes the fractile graph from the sample with index i .

$$\text{Let} \quad \begin{aligned} &(y_1^1, x_1^1), \dots (y_n^1, x_n^1) \\ &(y_1^2, x_1^2), \dots (y_n^2, x_n^2) \end{aligned}$$

be two independent samples from the same bivariate population P^{12} . The pooled sample is $(y_1^{12}, x_1^{12}), \dots (y_{2n}^{12}, x_{2n}^{12})$. The sequences $(v_1^1, \dots v_g^1)$, $(v_1^2, \dots v_g^2)$ and $(v_1^{12}, \dots v_g^{12})$ are easily defined. (Instead of m we put $2m$ in the last case because the sample is of size $2n$).

$$\text{Let} \quad \begin{aligned} &(y_1^3, x_1^3) \dots (y_n^3, x_n^3) \\ &(y_1^4, x_1^4) \dots (y_n^4, x_n^4) \end{aligned}$$

be two independent samples from another bivariate population P^{34} . Samples (3) and (4) pooled together yield $(y_1^{34}, x_1^{34}), \dots (y_{2n}^{34}, x_{2n}^{34})$. The sequences $(v_1^3, \dots v_g^3)$, $(v_1^4, \dots v_g^4)$, $(v_1^{34}, \dots v_g^{34})$ are likewise defined.

2.3. The error area A_{12} for the two samples (1) and (2), is defined as the area between G^1 and G^2 bounded by the ordinates at 1 and g . A_{34} is the error area between samples (3) and (4). The separation A_* is defined as the error area between the samples (12) and (34). We now give algebraic expression for the error areas.

$$\begin{aligned} \text{Let} \quad &v_i^1 - v_i^2 = w_{i(12)} \\ &v_i^3 - v_i^4 = w_{i(34)}; \quad i = 1, \dots g \quad \dots (2.3.1) \\ &v_i^{12} - v_i^{34} = w_{i(*)}. \end{aligned}$$

The area between G^1 and G^2 between the ordinates at i and $i+1$ (we will omit the subscript (12) for typographical convenience)

$$\frac{1}{2}|w_i| + \frac{1}{2}|w_{i+1}| \quad \text{if } w_i w_{i+1} > 0,$$

i.e., if the curves do not intersect between i and $i+1$

$$\frac{1}{2}|w_i| + \frac{1}{2}|w_{i+1}| - \frac{|w_i w_{i+1}|}{|w_i| + |w_{i+1}|} \quad \text{if } w_i w_{i+1} < 0 \quad \dots (2.3.2)$$

i.e., if the curves intersect between i and $i+1$.

Thus

$$A_{12} = \sum_{i=1}^{g-1} \left\{ \frac{1}{2}|w_i| + \frac{1}{2}|w_{i+1}| - \partial(w_i, w_{i+1}) \frac{|w_i w_{i+1}|}{|w_i| + |w_{i+1}|} \right\} \quad \dots (2.3.3)$$

where

$$\partial(a, b) = \begin{cases} 0 & \text{if } ab > 0 \\ 1 & \text{if } ab < 0 \end{cases}$$

LIMIT DISTRIBUTIONS CONNECTED WITH FRACTILE GRAPHICAL ANALYSIS

2.4. A new statistic Δ which may be called the norm statistic and which may be used instead of the error area is defined below. For the curves G^1 and G^2 the norm statistic Δ_{12} is given by

$$\Delta_{12} = \sqrt{w_{1(12)}^2 + \dots + w_{g(12)}^2} \quad \dots \quad (2.4.1)$$

Δ_{34} is the norm statistic between G^3 and G^4 . Δ_* is the norm statistic between G^{12} and G^{34} . Δ^2 is a particular case of a more general class of statistics that can be defined as soon as a positive definite matrix $B = (b_{ij})$ is given. Thus for the two fractile graphs G^1 and G^2 we define

$$\Gamma_{12}^2 = \sum_i \sum_j w_{i(12)} w_{j(12)} b_{ij}. \quad \dots \quad (2.4.2)$$

Similarly we define Γ_{34}^2 for G^3 and G^4 and Γ_*^2 for G^{12} and G^{34} .

2.5. A, Δ, Γ are all, in a way, measures of consistency between two samples. These statistics may also be used for testing in the two sample problem. The advantage associated with these statistics is that they provide a comparison between the two samples over the whole range. An inequality that connects the norm statistics Δ with error area A is

$$\frac{\Delta}{6} \leq A \leq \Delta \sqrt{g}. \quad \dots \quad (2.5.1)$$

2.6. The analysis can be suitably modified to yield concentration curves and concentration ratios when X and Y are positive, thus : plot the points $z_0 = 0$, $z_i = \frac{v_1 + \dots + v_i}{v_1 + \dots + v_g}$; $i = 1, \dots, g$, against the equidistant points $0, i/g, i = 1, \dots, g$ and join the points successively. We get a curve C known as specific concentration curve when the variables Y and X are different and as the Lorenz concentration curve when $Y = X$. Σ , the concentration ratio is defined as the twice the area between the curve C and the straight line joining the points $(0, 0)$ and $(1, 1)$. The algebraic expression for Σ is given by

$$\Sigma = \frac{2 \sum_{i=1}^g i v_i}{g \sum_{i=1}^g v_i} - \frac{g+1}{g}$$

3. SUMMARY OF RESULTS

3.1. (a) It is shown under some mild conditions that the asymptotic distribution of $m\Delta_{12}^2 = R_{12}$ is a mixture of χ^2 distributions, which we know can be approximated by a single χ^2 distribution.

(b) For a suitably chosen B , namely $\left[\frac{\Lambda^{-1}}{2}\right]$ defined in equation (5.3.8), $m\Gamma_{12}^2 = Q_{12}$ is asymptotically distributed as a χ^2 distribution with g degrees of freedom.

(c) It has been shown, in general, that $E(\Gamma^2) \simeq \frac{g}{m}$ and in case (Y, X) is distributed as a bivariate normal, that $E(\Delta^2) \simeq \frac{\text{constant} \times g}{m}$ where $\simeq$ means 'approximately equal to'. Takeuchi (1961) has shown that when (Y, X) is bivariate normal, $E(A) \simeq \frac{\text{constant} \times g}{\sqrt{m}}$. It is also shown that, if g is fixed then $E(\Delta^2)$ and $E(\Gamma^2)$ tend to zero at the rate of $\frac{1}{m}$, and from inequality (2.5.1) that $E(A)$ tends to zero at the rate of $\frac{1}{\sqrt{m}}$.

(d) We have seen that $m \Delta_{12}^2 = R_{12}$ is asymptotically distributed as a mixture of χ^2 distributions. It is shown that R_{12} and $R_* = 2m\Delta_*^2$ are asymptotically independent so that

$$\frac{R_*}{R_{12} + R_{34}} = \frac{2\Delta_*^2}{\Delta_{12}^2 + \Delta_{34}^2}$$

is distributed asymptotically as the ratio of independent mixtures of χ^2 distributions. This distribution can be approximated by an F distribution. Also Q_{12} and $Q_* = 2m\Gamma_*^2$ are shown to be asymptotically independent so that

$$\frac{Q_*}{Q_{12} + Q_{34}} = \frac{2\Gamma_*^2}{\Gamma_{12}^2 + \Gamma_{34}^2}$$

is distributed asymptotically as an F distribution with g and $2g$ degrees of freedom. We will incidentally derive that the combined fractile graph G^{12} tends to lie wholly within the two fractile graphs G^1 and G^2 as $m \rightarrow \infty$.

(e) Let r be the number of intersections between the curves G^1 and G^2 . r is a random variable and may take any value from 0 to $g-1$. First it was surmised that r behaves as the number of changes in the 'runs' of heads in independent tosses of an unbiased coin. It is proved here that this is not true. The asymptotic mean of r is derived.

(f) It is shown that the concentration ratio Σ is asymptotically normal.

4. A USEFUL THEOREM

4.1. In many cases when we are interested in the asymptotic distribution of a pair of random variables (Y_n, X_n) we are confronted with a situation like this: The sequence of conditional distributions $F_n(y/x)$ (assumed to exist for all x for each n) converge in law to a distribution $F(y/x)$ each x . Also the marginal distributions $G_n(x)$ of X_n converge in law to $G(x)$. Does it follow that the joint distribution of (Y_n, X_n) namely

$$H_n(y, x) \text{ converges in law to } \int_{-\infty}^x F_n(y/t) dG(t)? \quad \dots (4.1.1)$$

LIMIT DISTRIBUTIONS CONNECTED WITH FRACTILE GRAPHICAL ANALYSIS

Theorem 1: *The assertion contained in equation (4.1.1) is true if the convergence of the marginal distributions is slightly stronger, namely, the sequence of measures $\mu_n(E)$ corresponding to $G_n(x)$ converges on every Borel set E to the measure $\mu(E)$ corresponding to $G(x)$.*

The proof of this theorem and other related theorems under investigation will be published elsewhere.

5. SOME LIMIT DISTRIBUTIONS

5.1. First we develop some more notations.

$F(y, x)$ is the distribution function of a random variable (Y, X) . We assume that this distribution admits of a density function $f(y, x)$ which is continuous. The means μ and ν of X and Y respectively are finite. σ_X^2 , $\rho\sigma_X\sigma_Y$, σ_Y^2 the variances and covariance of X and Y are also finite.

The marginal distribution function of X is $G(x)$ with a density function $g(x)$ which is continuous. $\theta_0, \theta_1, \dots, \theta_{g-1}, \theta_g$ are defined by the relations

$$G(\theta_i) = \frac{i}{g}; \quad i = 1, \dots, g-1, \\ \theta_0 = -\infty, \theta_g = +\infty. \quad \dots \quad (5.1.1)$$

We define $g(\theta_0) = 0, g(\theta_g) = 0$ and assume that $g(\theta_i) \neq 0, i = 1, g-1$.

We next define the truncated means, variances and covariances as follows

$$\mu_i = g \int_{\theta_{i-1}}^{\theta_i} x g(x) dx; \quad i = 1, \dots, g \quad \dots \quad (5.1.2)$$

$$\nu_i = g \int_{\theta_{i-1}}^{\theta_i} \int_{-\infty}^{\infty} y f(y, x) dx dy; \quad i = 1, \dots, g \quad \dots \quad (5.1.3)$$

$$\sigma_i^2 = g \int_{\theta_{i-1}}^{\theta_i} (x - \mu_i)^2 g(x) dx; \quad i = 1, \dots, g \quad \dots \quad (5.1.4)$$

$$\tau_i^2 = g \int_{\theta_{i-1}}^{\theta_i} \int_{-\infty}^{\infty} (y - \nu_i)^2 f(y, x) dx dy; \quad i = 1, \dots, g \quad \dots \quad (5.1.5)$$

$$\rho_i \sigma_i \tau_i = g \int_{\theta_{i-1}}^{\theta_i} \int_{-\infty}^{\infty} (x - \mu_i)(y - \nu_i) f(y, x) dx dy; \quad i = 1, \dots, g. \quad \dots \quad (5.1.6)$$

We define some more quantities

$$\lambda(\theta_i) = E(Y|X = \theta_i); \quad i = 1, \dots, g-1 \quad \dots \quad (5.1.7) \\ \lambda(\theta_0) = \lambda(\theta_g) = 0.$$

5.2. Let $(y_1, x_1), \dots, (y_n, x_n)$ be n independent observations on the random variable (Y, X) and rearrange the sample thus,

$$(y_{(1)}, x_{(1)}), \dots, (y_{(n)}, x_{(n)})$$

so that

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

Assuming that $n = m.g$ where m and g are integers, we define

$$u_i = \frac{1}{m} \sum_{r=(i-1)m+1}^{im} x_{(r)}, \quad v_i = \frac{1}{m} \sum_{r=(i-1)m+1}^{im} y_{(r)} \quad i = 1, \dots, g \quad \dots (5.2.1)$$

$$\left. \begin{aligned} \sqrt{m}(u_i - \mu_i) &= \xi_i(n) \\ \sqrt{m}(v_i - \nu_i) &= \eta_i(n) \\ \sqrt{n}(x_{(im+1)} - \theta_i) &= \zeta_i(n); \quad i = 1, \dots, g-1 \\ \zeta_0(n) &= \zeta_g(n) = 0. \end{aligned} \right\} \quad \dots (5.2.2)$$

Also, let $\bar{x}$ and $\bar{y}$ be the sample means,

$$\begin{aligned} \sqrt{n}(\bar{x} - \mu) &= \xi(n) \\ \sqrt{n}(\bar{y} - \nu) &= \eta(n). \end{aligned} \quad \dots (5.2.3)$$

5.3. The main theorem of this section is

Theorem 2 : *The distribution of $\eta(n) = (\eta_1(n), \dots, \eta_g(n))$ tends to a multivariate normal distribution.*

Proof: To prove this theorem we proceed as follows :

We know that the distribution of $\xi(n) = (\xi_1(n), \dots, \xi_{g-1}(n))$ tends to a multivariate normal distribution N_1 in the strong sense of Theorem 1 since the densities converge. For instance see Cramer (1946). We will demonstrate the

Lemma 1 : *The conditional distribution of $(\xi(n), \eta(n)) = (\xi_1(n), \dots, \xi_g(n), \eta_1(n), \dots, \eta_g(n))$ given $\xi(n) = \xi$ tends to a multivariate normal distribution where the asymptotic means of $(\xi(n), \eta(n))$ are linear functions of the ξ 's and the asymptotic variances and covariances are independent of ξ .*

An application of Theorem 1 shows that the distribution of $(\xi(n), \eta(n), \zeta(n))$ tends to a multivariate normal distribution. In particular the distribution of $\eta(n)$ tends to a multivariate normal distribution. Hence the theorem.

Proof of Lemma 1 : We prove the lemma for the case $g = 2$.

$$\left. \begin{aligned} \xi_2^0(n) &= \sqrt{m} \left(\frac{x_{(m+2)} + \dots + x_{(2m)}}{m} - \mu_2 \right) \\ \eta_2^0(n) &= \sqrt{m} \left(\frac{y_{(m+2)} + \dots + y_{(2m)}}{m} - \nu_2 \right). \end{aligned} \right\} \quad \dots (5.3.1)$$

LIMIT DISTRIBUTIONS CONNECTED WITH FRACTILE GRAPHICAL ANALYSIS

As soon as $\zeta_1(n)$ is fixed at ζ_1 the sample splits up into three independent parts—namely:

m independent observations on the random variable (Y, X) truncated to the region

$$-\infty < X < \theta_1 + \frac{\zeta_1}{\sqrt{n}}$$

$(m-1)$ independent observations on the random variable (Y, X) truncated to the region

$$\theta_1 + \frac{\zeta_1}{\sqrt{n}} < X < \infty,$$

and one observation at

$$\theta_1 + \frac{\zeta_1}{\sqrt{n}}.$$

Thus $(\xi_1(n), \eta_1(n))$ and $(\xi_2^0(n), \eta_2^0(n))$ are the normalised means of the samples in the two independent portions. In virtue of the Central Limit Theorem the distribution of these random variables given $\zeta_1(n) = \zeta_1$ tends to a multivariate normal distribution. The asymptotic means and variances are readily calculated. Let us denote the asymptotic mean and variance of a sequence of random variables, Z_n by $\bar{E}(Z_n)$ and $\bar{V}(Z_n)$. We find

$$\begin{aligned} \bar{E}(\xi_1(n) | \zeta_1(n) = \zeta_1) &= \sqrt{2}(\theta_1 - \mu_1)g(\theta_1)\zeta_1 \\ \bar{E}(\eta_1(n) | \zeta_1(n) = \zeta_1) &= \sqrt{2}(\lambda(\theta_1) - \nu_1)g(\theta_1)\zeta_1 \\ \bar{V}(\xi_1(n) | \zeta_1(n) = \zeta_1) &= \sigma_1^2 \\ \bar{V}(\eta_1(n) | \zeta_1(n) = \zeta_1) &= \tau_1^2 \\ \text{cov}(\xi_1(n), \eta_1(n) | \zeta_1(n) = \zeta_1) &= \rho_1\sigma_1\tau_1. \end{aligned} \quad \dots (5.3.2)$$

Similar expressions for $(\xi_2^0(n), \eta_2^0(n))$ hold.

We next show that $(\xi_2(n), \eta_2(n))$ and $(\xi_2^0(n), \eta_2^0(n))$ are asymptotically equivalent, i.e., $E(\xi_2(n) - \xi_2^0(n))^2 + E(\eta_2(n) - \eta_2^0(n))^2$ tends to zero as $n \rightarrow \infty$. We have to show that

$$\frac{1}{m} E(x_{(m+1)})^2 + \frac{1}{m} E(y_{(m+1)})^2 \rightarrow 0 \quad \dots (5.3.3)$$

$$\begin{aligned} E(x_{(m+1)}^2) &= E_{\zeta_1} E(x_{(m+1)}^2 | \zeta_1(n) = \zeta_1) \\ &= E \left(\theta_1 + \frac{\zeta_1}{\sqrt{n}} \right)^2 < \infty \end{aligned} \quad \dots (5.3.4)$$

$$\begin{aligned} E(y_{(m+1)}^2) &= E_{\zeta_1} E(y_{(m+1)}^2 | \zeta_1(n) = \zeta_1) \\ &= E \left\{ \frac{1}{g \left(\theta_1 + \frac{\zeta_1}{\sqrt{n}} \right)} \int_{-\infty}^{\infty} y^2 f \left(\theta_1 + \frac{\zeta_1}{\sqrt{n}}, y \right) dy \right\} < \infty \end{aligned} \quad \dots (5.3.5)$$

(5.3.4) and (5.3.5) are enough to prove (5.3.3). Thus the lemma is proved when $g = 2$.

A similar proof holds for any g . In the case of g groups we have

$$\left. \begin{aligned} \bar{E}(\xi_i(n) | \boldsymbol{\zeta}(n) = \boldsymbol{\zeta}) &= \sqrt{g}(\theta_i - \mu_i)g(\theta_i)\zeta_i - \sqrt{g}(\theta_{i-1} - \mu_i)g(\theta_{i-1})\zeta_{i-1} \\ \bar{E}(\eta_i(n) | \boldsymbol{\zeta}(n) = \boldsymbol{\zeta}) &= \sqrt{g}(\lambda(\theta_i) - \nu_i)g(\theta_i)\zeta_i - \sqrt{g}(\lambda(\theta_{i-1}) - \nu_i)g(\theta_{i-1})\zeta_{i-1} \end{aligned} \right\} \dots \quad (5.3.6)$$

$$\bar{V}(\xi_i(n) | \boldsymbol{\zeta}(n) = \boldsymbol{\zeta}) = \sigma_i^2$$

$$\bar{V}(\eta_i(n) | \boldsymbol{\zeta}(n) = \boldsymbol{\zeta}) = \tau_i^2$$

$$\overline{\text{cov}}(\xi_i(n), \eta_i(n) | \boldsymbol{\zeta}(n) = \boldsymbol{\zeta}) = \rho_i \sigma_i \tau_i \quad \dots \quad (5.3.7)$$

$$\overline{\text{cov}}(\xi_i(n), \xi_j(n) | \boldsymbol{\zeta}(n) = \boldsymbol{\zeta})$$

$$= \overline{\text{cov}}(\xi_i(n), \eta_j(n) | \boldsymbol{\zeta}(n) = \boldsymbol{\zeta})$$

$$= \overline{\text{cov}}(\eta_i(n), \eta_j(n) | \boldsymbol{\zeta}(n) = \boldsymbol{\zeta}) = 0, \quad i \neq j.$$

Let us denote an equation in (5.3.6) by

$$\bar{E}(\boldsymbol{\eta}(n) | \boldsymbol{\zeta}(n) = \boldsymbol{\zeta}) = \boldsymbol{\zeta} M \quad \text{where } M \text{ is a } g-1 \times g \text{ matrix.}$$

Let

$$S = \begin{pmatrix} \tau_1^2 & 0 & \dots \\ 0 & \tau_2^2 & \dots \\ \dots & \dots & \dots \\ \dots & \dots & \tau_g^2 \end{pmatrix}$$

Let us denote the variance covariance matrix of $\boldsymbol{\zeta}(n)$ in the limit by

$$K = (k_{ij}).$$

$$k_{ij} = \frac{i(g-j)}{g^2} \cdot \frac{1}{g(\theta_i)g(\theta_j)} \quad i \leq j, \quad i, j = 1, \dots, g-1.$$

Then the asymptotic distribution of $\boldsymbol{\eta}(n)$ has for its density function

$$\text{const. exp} - \left(\frac{\boldsymbol{\eta} \Lambda^{-1} \boldsymbol{\eta}'}{2} \right)$$

where

$$\Lambda = (\lambda_{ij}) = S + M' K M. \quad \dots \quad (5.3.8)$$

Define

$$M_i = i(\lambda(\theta_i) - \nu_i) - (i-1)(\lambda(\theta_{i-1}) - \nu_i)$$

$$M'_i = (g-i)(\lambda(\theta_i) - \nu_i) - (g-i+1)(\lambda(\theta_{i-1}) - \nu_i) \quad \dots \quad (5.3.9)$$

LIMIT DISTRIBUTIONS CONNECTED WITH FRACTILE GRAPHICAL ANALYSIS

Using (5.3.8) and (5.3.9) we can write λ_{ij} as

$$\lambda_{ij} = \frac{1}{g} M_i M'_j \quad \text{if } j > i$$

$$= \frac{1}{g} M_i M'_i + (\lambda(\theta_{i-1}) - v_i)(\lambda(\theta_i) - v_i) + \tau_i^2 \quad \text{if } i = j$$

Let us denote the correlation coefficient between η_i and η_{i+1} by α_i , $i = 1, \dots, g-1$. If the regression of Y on X monotonic, we can easily see that all the α_i 's are positive.

6. APPLICATION OF THE RESULTS IN SECTION 5

6.1. In Section 2 we saw that our statistics A , Δ , Γ were all function of $w_{(12)} \dots$ etc.

Define

$$\psi_{i(12)}(n) = \sqrt{m} w_{i(12)} = \sqrt{m(v_i^1 - v_i^2)}$$

$$= \eta_{i(1)}(n) - \eta_{i(2)}(n). \quad \dots \quad (6.1.1)$$

Theorem 2 shows that the distribution of $(\psi_{(12)}(n)) = (\psi_{1(12)}(n), \dots, \psi_{g(12)}(n))$ tends to a multivariate normal distribution with mean zero and variance covariance matrix 2Λ . Hence $R_{12} = m\Delta_{12}^2 = m\Sigma w_{i(12)}^2 = \Sigma \psi_{i(12)}^2(n)$ has an asymptotic distribution which is a mixture of χ^2 distributions, the compounding coefficients being the latent roots $\lambda_1, \dots, \lambda_g$ of the matrix 2Λ . Under the assumption that the populations P^{12} and P^{34} are identical we see that $m\Delta_{34}^2 = R_{34}$ and $2m\Delta_*^2 = R_*$ have the same asymptotic distribution. Taking $B = \frac{\Lambda^{-1}}{2}$ in equation (2.4.2) we get

$$\Gamma_{12} = \mathbf{w}_{(12)} \left(\frac{\Lambda^{-1}}{2} \right) \mathbf{w}'_{(12)}$$

i.e.

$$Q_{12} = m\Gamma_{12}^2 = \psi_{(12)}(n) \left(\frac{\Lambda^{-1}}{2} \right) \psi_{(12)}(n)$$

and this is distributed asymptotically as a χ^2 with g degrees of freedom. Similarly $Q_{34} = m\Gamma_{34}^2$ and $Q_* = 2m\Gamma_*^2$ have the same asymptotic distribution.

6.2. We also see that

$$\bar{E}(\Delta^2) = \frac{\lambda_1 + \dots + \lambda_g}{m} = \frac{\sum_1^g \lambda_{ii}}{m}$$

$$\bar{E}(\Gamma^2) = \frac{g}{m} \quad \dots \quad (6.2.1)$$

From Takeuchi (1961) we see, that if (Y, X) is distributed as a bivariate normal with parameters $\mu_2, \mu_1, \sigma_2^2, \sigma_1^2, \rho$, then $\frac{1}{g} \sum_1^g \lambda_{ii} \rightarrow \sigma_2^2 (1 - \rho^2)$ as $g \rightarrow \infty$.

Thus $\bar{E}(\Delta^2) \rightarrow \frac{\text{constant} \times g}{m}$ as $g \rightarrow \infty$.

The results about the expectations stated in the summary follow.

6.3. We shall later on prove the following.

Lemma 2 : $\psi_{i(12)}(n)$ and $\psi_{j(*)}(n)$ are asymptotically independent for any i and j .

From this lemma it follows that R_{12} and R_* are asymptotically independent so that

$$\frac{R_*}{R_{12} + R_{34}} = \frac{2\Delta_*^2}{\Delta_{12}^2 + \Delta_{34}^2}$$

is asymptotically distributed as the ratio of independent mixtures of distributions. It also follows that Q_* and Q_{12} are asymptotically independent. Hence

$\frac{Q_*}{Q_{12} + Q_{34}} = \frac{2\Gamma_*^2}{\Gamma_{12}^2 + \Gamma_{34}^2}$ is asymptotically distributed as an F with g and $2g$ degrees of freedom.

Proof of Lemma 2 :

$$\left. \begin{aligned} \psi_{i(12)}(n) &= \sqrt{m}(v_i^1 - v_i^2) = \eta_{i(1)}(n) - \eta_{i(2)}(n) \\ \psi_{j(*)}(n) &= \sqrt{2m}(v_j^{12} - v_j^{34}) \end{aligned} \right\} \quad \dots \quad (6.3.1)$$

Suppose we show that $\sqrt{m} \left(v_j^{12} - \frac{v_j^1 + v_j^2}{2} \right) \rightarrow 0$ in probability, i.e., $\psi_{j(*)}(n)$ is asymptotically equivalent to

$$\begin{aligned} &\sqrt{2} \sqrt{m} (v_j^1 + v_j^2 - v_j^3 - v_j^4) \\ &= \sqrt{2}(\eta_{j(1)}(n) + \eta_{j(2)}(n) - \eta_{j(3)}(n) - \eta_{j(4)}(n)). \end{aligned} \quad \dots \quad (6.3.2)$$

From the expressions (6.3.1) and (6.3.2) we at once deduce that $\psi_{i(12)}(n)$ and $\psi_{j(*)}(n)$ are asymptotically distributed as a bivariate distribution normal with correlation zero, and hence are asymptotically independent.

It now remains to show that

$$\alpha_j(n) = \sqrt{m} \left(v_j^{12} - \frac{v_j^1 + v_j^2}{2} \right) \rightarrow 0 \text{ in probability.}$$

or that

$$E(\alpha_j^2(n)) \rightarrow 0 \text{ as } n \rightarrow \infty$$

We assume that $g = 2$.

Then $\alpha_1(n) = (\Sigma_1 y_{(r)}^1 - \Sigma_2 y_{(r)}^2) \sqrt{m}$ in case $x_{(m+1)}^1 \leq x_{(m+1)}^2$

where Σ_1 is the summation over all indices r such that $x_{(m+1)}^1 \leq x_{(r)}^1 < x_{(2m+1)}^{12}$ and Σ_2 is the summation over all indices such that $x_{(2m+1)}^{12} \leq x_{(r)}^2 < x_{(m+1)}^2$. We have another similar expression in case $x_{(m+1)}^2 < x_{(m+1)}^1$. The probability of either event is $\frac{1}{2}$. From these expressions for $\alpha_1(n)$ we see that to prove $E(\alpha_1^2(n)) \rightarrow 0$ it is enough

to prove that $E\left(\frac{1}{\sqrt{m}} \sum_1 y_{(r)}^1\right)^2 \rightarrow 0$ as $n \rightarrow \infty$. Let s be the number of observations included in the summation Σ_1 . Naturally s is a random variable. We give the following expressions that can be derived after some algebra. Let us denote $\sqrt{n}(x_{(2m+1)}^{12} - \theta_1)$ by ξ_1^0 .

$$E\left[\left(\frac{1}{\sqrt{m}}(\Sigma_1 y_{(r)}^1 - \Sigma_2 y_{(r)}^2)\right)^2 \mid \xi_1^0, \xi_1^2, \xi_1^0, s, \xi_1^1 < \xi_1^0\right] = \frac{s}{m} \times \text{const} \times (1 + o(1)) \quad \dots (6.3.3)$$

$$E\left(\frac{s}{m} \mid \xi_1^1, \xi_1^0, \xi_1^1 < \xi_1^0\right) = \frac{\xi_1^0 - \xi_1^1}{\sqrt{m}} + o\left(\frac{|\xi_1^0 - \xi_1^1|}{\sqrt{m}}\right) \quad \dots (6.3.4)$$

$$E(|\xi_1^0 - \xi_1^1| \mid \xi_1^1 < \xi_1^0) < \infty. \quad \dots (6.3.5)$$

Expressions (6.3.3) through (6.3.5) suffice to show that $\alpha_1(n) \rightarrow 0$ as $n \rightarrow \infty$. The case of general g is dealt with in a similar way.

$\alpha_j(n) = \sqrt{m} \left(v_j^{12} - \frac{v_i^1 + v_j^2}{2}\right)$ tends to zero in probability means that the ordinate of the combined fractile graph G^{12} tends to the mean of the ordinates of the fractile graphs G^1 and G^2 . Thus we see that the graph G^{12} tends to lie completely within the graph G^1 and G^2 .

6.4. Let r be the number of intersections between the fractile graphs G^1 and G^2 . r is a random variable. We know that these graphs intersect between i and $i+1$ if and only if $w_{i(12)} w_{i+1(12)} < 0$, i.e., $\eta_{i(12)}(n) \eta_{i+1(12)} < 0$, i.e., the sign of $\eta_{i(12)}(n)$ and $\eta_{i+1(12)}(n)$ are opposite. Let us replace the sequence $(\eta_{1(12)}(n), \dots, \eta_{g(12)}(n))$ by their signs thus $(+, -, \dots, -)$ (say). The number of changes of 'runs' in this sequence, i.e., the total number of runs in this sequence minus one, is the number of intersections between the graphs G^1 and G^2 . The probability that $\eta_{i(12)}(n) < 0$ is $\frac{1}{2}$ for any i . But $\eta_{i(12)}(n)$ and $\eta_{j(12)}(n)$ are not independent even asymptotically for any distribution. This feature mars the beauty of the picture and the theory of runs with independent tosses of a coin cannot be applied here.

Anyhow we will calculate $E(r)$. Define the random variables $Z_i, i = 1, \dots, g-1$ by the relations

$$Z_i = \begin{cases} 1 & \text{if } \eta_i \eta_{i+1} < 0 \\ 0 & \text{if } \eta_i \eta_{i+1} \geq 0 \end{cases}$$

Then $r = \sum_{i=1}^{g-1} Z_i$, from this we find

$$\overline{E}(r) = \sum_{i=1}^{g-1} \frac{\cos^{-1} \alpha_i}{\pi} \quad \dots (6.4.1)$$

We note an interesting fact that $E(r) \leq \frac{g-1}{2}$ if all the α_i 's are positive.

6.5. The concentration ratio Σ is $\frac{2 \sum_1^g i v_i}{g \sum_1^g v_i} - \frac{g+1}{2g}$ (see (2.6.1)). From

Theorem 4.1 we see that the v_i 's are asymptotically distributed as a multivariate normal. Hence $\sqrt{n}(\Sigma - \Theta)$ is asymptotically distributed with mean zero and variance $\Lambda \Lambda'$ where $\mathbf{1} = (l_1 \dots l_g)$

$$l_i = \frac{2 \sum_{r=1}^g (i-r) v_i}{g \left(\sum_{r=1}^g v_i \right)^2}, \quad \Theta = \frac{2 \sum_1^g i v_i}{g \sum_1^g v_i} - \frac{g+1}{2}$$

7. REMARKS AND SUGGESTIONS

7.1. We have seen that if two samples (1) and (2) are taken from the same population then $\pi = E\eta_{(12)}(n) = 0$. Thus $\pi = 0$ may form the hypothesis for consistency of the samples. Since we know that $\eta(n)$ is asymptotically normally distributed, and we have only one set of observations namely $\eta_{(12)}(n)$, the best test, in a sense which can be made precise large samples, would be that based on

$$\eta_{(12)}(n) \left(\frac{\Lambda^{-1}}{2} \right) \eta_{(12)}(n) = \Gamma_{12}^2.$$

Starting from Theorem 5.1, the study of the distribution of the error area A should be possible and would be interesting. Again if Λ were estimated in practice the change in the asymptotic distribution of Γ^2 has to be studied.

8. ACKNOWLEDGEMENTS

I wish to thank Professor C. R. Rao and P. K. Bhattacharya of the Indian Statistical Institute for the useful discussions and criticisms.

REFERENCES

- CRAMÉR, (1946) : *Mathematical Methods of Statistics*, Princeton University Press.
 MAHALANOBIS, P. C. (1958): Lectures in Japan : Fractile Graphical Analysis : Indian Statistical Institute.
 ——— (1963): A method of Fractile Graphical Analysis. *Econometrica*, 28, 2, 325-351, reprinted in *Sankhyā*, Series A, 23, 41-64.
 TAKEUCHI, K. (1961): On some properties of error area in the fractile graph method. *Sankhyā*, Series A, 23, 65-78.

Paper received : November, 1958.

SOME LIMIT THEOREMS IN REGRESSION THEORY

By K. R. PARTHASARATHY

and

P. K. BHATTACHARYA

Indian Statistical Institute

SUMMARY. (X, Y) follows an unknown bivariate distribution with $0 \leq X \leq 1$ and the regression of Y on X is continuous and a sequence of observations on (X, Y) are made. An estimate of the unknown regression function based on these observations and motivated by the method of Fractile Graphical Analysis has been suggested. Its large sample properties, viz., convergence in probability and almost sure uniform convergence to the true regression function have been investigated. Large sample tests for a specified regression function have also been proposed for the case when the conditional variance function of Y on X is known and for the case when it is unknown.

1. INTRODUCTION

Let us suppose that X and Y are real valued variables having a certain joint distribution function such that X takes values in the interval $[0, 1]$ and all conditional absolute moments of order up to $p \geq 3$ exist when X is fixed at any point x . If we can make a sequence of independent observations on (X, Y) the question naturally arises as to how we can construct from these observations an estimate of the unknown regression function of Y on X , possessing certain properties like convergence in probability, almost sure uniform convergence etc., to the true regression function. Another important problem is that of constructing at least a large sample test for a specified regression function. In Section 2 we shall make use of the technique of Fractile Graphical Analysis suggested by Mahalanobis (1958) to estimate the regression and analyse its large sample properties and in Section 3, construct a large sample test for the regression. The crucial part of our analysis consists of the utilization of certain results concerning the error of approximation by the central limit theorem and an upper bound of the tail probabilities in the distribution of sums of independent and bounded random variables. These results are given in the Appendix for reference.

2. ESTIMATION OF THE REGRESSION FUNCTION

Let $(x_1, y_1), (x_2, y_2), \dots, (x_{kn}, y_{kn})$ be independent observations on (X, Y) , $x_{(r)}$ the r -th order statistic in the set of observed values of X and $y_{(i)} = y_j$ if $x_{(i)} = x_j$. Throughout this section we shall assume that the distribution function of X is continuous and strictly increasing so that the probability of any two observations on X being

equal is zero. Then the statistics $x_{(r)}$ are defined unambiguously for almost all samples. In order to simplify notation we write for $r = 1, 2, \dots, k$ and $s = 1, 2, \dots, n$,

$$\begin{aligned} x_{(r-1)n+s} &= x_{rs}, & y_{(r-1)n+s} &= y_{rs}, \\ E[Y|X=x] &= \varphi(x), & E[Y^2|X=x] &= \xi(x), \\ E[|Y-\varphi(x)|^m|X=x] &= \beta_m(x), & m &= 2, 3, \dots, p, \\ \varphi(x_{rs}) &= \varphi_{rs}, & \xi(x_{rs}) &= \xi_{rs}, & \beta_m(x_{rs}) &= \beta_{mrs}, \\ \bar{y}_r &= \sum_{s=1}^n y_{rs}/n, & \bar{y}_r^2 &= \sum_{s=1}^n y_{rs}^2/n, \\ \bar{\varphi}_r &= \sum_{s=1}^n \varphi_{rs}/n, & \bar{\varphi}_r^2 &= \sum_{s=1}^n \varphi_{rs}^2/n, \\ \bar{\beta}_{mr} &= \sum_{s=1}^n \beta_{mrs}/n, & \bar{\xi}_r &= \sum_{s=1}^n \xi_{rs}/n. \end{aligned} \quad \dots (2.1)$$

Now we define two functions $f_{nk}(x)$ and $\hat{\varphi}_{nk}(x)$ as follows :

$$\begin{aligned} f_{nk}(x) &= \bar{y}_1 & \text{if } 0 \leq x \leq x_{1n} \\ &= \bar{y}_r & \text{if } x_{r-1n} < x \leq x_{rn}, \quad r = 2, \dots, k-1 \\ &= \bar{y}_k & \text{if } x_{k-1n} < x \leq 1. \end{aligned} \quad \dots (2.2)$$

$$\begin{aligned} \varphi_{nk}(x) &= \bar{\varphi}_1 & \text{if } 0 \leq x \leq x_{1n} \\ &= \bar{\varphi}_r & \text{if } x_{r-1n} < x \leq x_{rn}, \quad r = 2 \dots k-1 \\ &= \bar{\varphi}_k & \text{if } x_{k-1n} < x \leq 1. \end{aligned} \quad \dots (2.3)$$

Lemma 1 : If the random variable X has a strictly increasing continuous distribution function $F(x)$, then

$$\begin{aligned} P \left[F^{-1} \left(\frac{r}{k} - \frac{1}{2k} \right) \leq x_{rn} \leq F^{-1} \left(\frac{r}{k} + \frac{1}{2k} \right), r = 1, 2, \dots, k-1 \right] \\ \geq 1 - 4m \cdot \exp \left[-\frac{3n}{16k} \right] - 2k \cdot \exp \left[-\frac{n}{4k} \left(1 - \frac{1}{8m} \right) \right] \end{aligned}$$

for every fixed positive integer m .

Proof : We have

$$P[x_{rn} < z] = \sum_{s=0}^{kn-rn} \binom{kn}{s} [1-F(z)]^s [F(z)]^{kn-s}. \quad \dots (2.4)$$

Applying Theorem (A 2) to the sum S_N of N independent binomial random variables each with probability for success equal to p , we have

$$P[S_N \geq N(p+\delta)] \leq \begin{cases} \exp \left[\frac{-N\delta}{2p} \log \left(\frac{\delta}{2q} + \sqrt{1 + \frac{\delta^2}{4q^2}} \right) \right] & \text{if } p \geq q \\ \exp \left[\frac{-N\delta}{2q} \log \left(\frac{\delta}{2p} + \sqrt{1 + \frac{\delta^2}{4p^2}} \right) \right] & \text{if } p \leq q, \end{cases}$$

which, if $\frac{\delta}{2p}$ and $\frac{\delta}{2q}$ are less than unity, becomes

$$P[S_N \geq N(p+\delta)] \leq \begin{cases} \exp \left[-N\delta^2 \left(1 - \frac{\delta}{4q} \right) \right] & \text{if } p \geq q \\ \exp \left[-N\delta^2 \left(1 - \frac{\delta}{4p} \right) \right] & \text{if } p \leq q. \end{cases} \quad \dots (2.5)$$

We have

$$\begin{aligned} P \left[F^{-1} \left(\frac{r}{k} - \frac{1}{2k} \right) \leq x_m \leq F^{-1} \left(\frac{r}{k} + \frac{1}{2k} \right), r = 1, 2, \dots, k-1 \right] \\ \geq 1 - \sum_{r=1}^{k-1} P \left[x_m < F^{-1} \left(\frac{r}{k} - \frac{1}{2k} \right) \right] - \sum_{r=1}^{k-1} P \left[x_m > F^{-1} \left(\frac{r}{k} + \frac{1}{2k} \right) \right] \end{aligned}$$

which, after using (2.4) and an application of (2.5) for $\delta = \frac{1}{2k}$ and two binomial sums with probability for successes $\frac{r}{k} - \frac{1}{2k}$ and $1 - \frac{r}{k} - \frac{1}{2k}$ and number of summands equal to nk , becomes greater than $4m \cdot \exp \left[-\frac{3n}{16k} \right] + 2k \cdot \exp \left[\frac{-n}{4k} \left(1 - \frac{1}{8m} \right) \right]$ for every fixed integer m .

Lemma 2: Under the conditions of Lemma 1 if $\varphi(x)$, the regression function of Y on X is continuous then for any $\epsilon > 0$

$$P \left[\sup_{0 \leq x \leq 1} |\hat{\varphi}_{nk}(x) - \varphi(x)| > \epsilon \right] < c(\epsilon, \varphi) \left[m \exp \left[-\frac{3n}{16k} \right] + k \exp \left[-\frac{n}{4k} \left(1 - \frac{1}{8m} \right) \right] \right]$$

for every fixed integer m , and where $c(\epsilon, \varphi)$ is a constant which depends on ϵ and φ only.

Proof: Let $\delta[a, b] = \max_{a \leq x_1, x_2 \leq b} |\varphi(x_1) - \varphi(x_2)|$

$$\delta_k = \max \left\{ \delta \left[0, F^{-1} \left(\frac{1}{2k} \right) \right], \delta \left[F^{-1} \left(\frac{1}{k} - \frac{1}{2k} \right), F^{-1} \left(\frac{1}{k} + \frac{1}{2k} \right) \right], \dots, \delta \left[F^{-1} \left(1 - \frac{1}{2k} \right), 1 \right] \right\}.$$

$$\text{If } F^{-1} \left(\frac{r}{k} - \frac{1}{2k} \right) \leq x_m \leq F^{-1} \left(\frac{r}{k} + \frac{1}{2k} \right), r = 1, 2, \dots, k-1$$

$$\text{then } d_{nk}(\varphi) = \sup_{0 \leq x \leq 1} |\hat{\varphi}_{nk}(x) - \varphi(x)| \leq \delta_k. \quad \dots (2.6)$$

Thus

$$\begin{aligned} P[d_{nk}(\varphi) > \epsilon] &\leq \left[P[d_{nk} > \epsilon | F^{-1} \left(\frac{r-1}{k} \right) \leq x_m \leq F^{-1} \left(\frac{r+1}{k} \right), r = 1, 2, \dots, k-1] \right] \\ &+ 1 - P \left[F^{-1} \left(\frac{r-1}{k} \right) \leq x_m \leq F^{-1} \left(\frac{r+1}{k} \right), r = 1, 2, \dots, k-1 \right] \quad \dots (2.7) \end{aligned}$$

Because of (2.6) the first term on the right side of (2.7) becomes less than $P(\delta_k > \epsilon)$. Since F is continuous and strictly increasing F^{-1} is continuous. Since

$\varphi(x)$ is uniformly continuous we have $\delta_k < \epsilon$ for $k > k_0(\epsilon, \varphi)$. Thus the first term vanishes for $k > k_0(\epsilon, \varphi)$. The second term becomes less than $4m \exp \left[\frac{-3n}{16k} \right] + 2k \exp \left[\frac{-n}{4k} \left(1 - \frac{1}{8m} \right) \right]$ because of Lemma 1. Combining the two we get the required inequality.

Hereafter, for simplicity, we shall assume that $\beta_2(x) > c_1 > 0$ and $\beta_m(x) < c_2$ for all $m = 2, \dots, p$ and all x . However, from the proofs we can see that all our results hold good under more general conditions.

Lemma 3: If $\beta_2(x) > c_1 > 0$ and $\beta_m(x) < c_2$ for all $m = 1, 2, \dots, p$ and all x , then for any $\epsilon > 0$

$$P \left[\sup_{0 \leq x \leq 1} |f_{nk}(x) - \hat{\varphi}_{nk}(x)| > \epsilon \right] < C \cdot \frac{k}{n^{p-1}}$$

where C is a constant depending only on ϵ, p, c_1 and c_2 .

Proof: We have

$$P(\epsilon) = P \left[\sup_{0 \leq x \leq 1} |f_{nk}(x) - \hat{\varphi}_{nk}(x)| > \epsilon | x_1, \dots, x_{kn} \right]$$

$$= P \left[\sup_{r=1, 2, \dots, k} |\bar{y}_r - \bar{\varphi}_r| > \epsilon | x_1, \dots, x_{kn} \right]$$

$$\leq \sum_{r=1}^h P \left[\frac{|\bar{y}_r - \bar{\varphi}_r|}{\sqrt{\beta_{2r}/n}} > \frac{\epsilon \sqrt{n}}{\sqrt{c_2}} \right]. \quad \dots (2.8)$$

Making use of (A7) and noting that $T_{pn} < \frac{\sqrt{n}}{4}$ we have

$$\frac{\epsilon \sqrt{n}}{\sqrt{c_2}} > \sqrt{4(p-2) \log T_{pn}} \text{ for } n > n(\epsilon).$$

Hence by Theorem A1 we obtain

$$P(\epsilon) \leq 2k \Phi \left(\frac{-\epsilon \sqrt{n}}{\sqrt{c_2}} \right) \cdot \frac{\theta_p c_3}{1 + \left(\frac{\epsilon \sqrt{n}}{\sqrt{c_2}} \right)^p} \cdot \frac{k}{n^{\frac{p-2}{2}}} \quad \dots (2.9)$$

where $c_3 = \frac{4c_2^{3/p}}{c_1^{3/2}}$ and θ_p is a positive constant which depends only on p and

$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$. Since $\Phi(-x) \leq \frac{1}{\sqrt{2\pi} x} e^{-x^2/2}$, $x > 0$, (2.9) gives, for $n > n(\epsilon)$,

$$\begin{aligned} P(\epsilon) &\leq \frac{k}{\epsilon \sqrt{n/c_2}} e^{\frac{-\epsilon^2}{2c_2} \cdot n} + \theta_p \cdot \frac{c_2^{p/2}}{\epsilon^p} \cdot \frac{c_3 k}{n^{p-1}} \\ &= c_4(\epsilon, p, c_1, c_2) \left[\frac{k}{\sqrt{n}} e^{\frac{-\epsilon^2}{2c_2} \cdot n} + \frac{k}{n^{p-1}} \right] \end{aligned}$$

which completes the proof.

SOME LIMIT THEOREMS IN REGRESSION THEORY

Now we shall state and prove the main result of this section.

Theorem 1: *If the p -th conditional absolute moment of Y exists when X is fixed at any point x , $p \geq 3$, X has a strictly increasing continuous distribution function in $[0, 1]$, the regression $\phi(x)$ of Y on X is continuous and $\beta_2(x) > c_1 > 0$, $\beta_m(x) < c_2$ for all $m = 1, 2, \dots, p$ and for all x , then for $n \geq (4+\delta)k \log k$, $\delta > 0$, $d_{nk} = \sup_x |f_{nk}(x) - \phi(x)|$ converges to zero in probability as $k \rightarrow \infty$ and for $n \geq (8+\delta)k \log k$, d_{nk} converges almost surely to zero as $k \rightarrow \infty$.*

Proof: The first part of the theorem is an immediate consequence of Lemmas 2 and 3 and the second part follows from the same lemmas and Borel-Cantelli lemma.

3. LARGE SAMPLE TEST OF A SPECIFIED REGRESSION FUNCTION

In this section we shall consider the problem of testing the null hypothesis H_0 that the regression function $\phi(x)$ is equal to a specified function $\mu(x)$.

Consider the statistic

$$\tau_{nk} = \sup_{r=1, \dots, k} \frac{|\bar{y}_r - \bar{\mu}_r|}{\sqrt{\bar{\beta}_{2r}/n}},$$

where $\bar{y}_r$ and $\bar{\beta}_{2r}$ are as defined in (2.1) and $\bar{\mu}_r = \frac{\sum_{s=1}^n \mu(x_{rs})}{n}$. If we had known the probability distribution of τ_{nk} under the null hypothesis, then for any given level of significance $0 < \alpha < 1$ we could apply the following test:

Reject H_0 if and only if $\tau_{nk} > \tau_{nk}(\alpha)$

where

$$P[\tau_{nk} \leq \tau_{nk}(\alpha) | H_0] = \alpha.$$

Theorem 2 enables us to apply such a test at least in the large sample.

Theorem 2: *If $\beta_2(x) > c_1 > 0$ and $\beta_m(x) < c_2 < \infty$ for all x and $m = 1, 2, \dots, p$, $p \geq 3$ then we have for $n \geq (k \log k)^{2/p-2}$*

$$\lim_{k \rightarrow \infty} P[\tau_{nk} \leq \lambda_k | H_0] = \exp \left[-\frac{1}{\sqrt{\pi}} e^{-\theta} \right]$$

where

$$\lambda_k = \sqrt{2(\theta + \log k - \frac{1}{2} \log \log k)}.$$

Proof: Under H_0 , we have

$$P(\lambda_k) = P[\tau_{nk} \leq \lambda_k | x_1, \dots, x_{kn}]$$

$$\prod_{r=1}^k P \left[\frac{|\bar{y}_r - \bar{\phi}_r|}{\sqrt{\bar{\beta}_{2r}/n}} \leq \lambda_k | x_1, \dots, x_{kn} \right] = \prod_{r=1}^k [\Phi(\lambda_k) - \Phi(-\lambda_k) + e_{rk}]$$

where because of (A8) and the fact that $\lambda_k < \sqrt{4(p-2) \log T_{2n}}$ for $k \geq k_0(\theta)$ we get

$$|e_{rk}| \leq c_5 \left[\frac{(1 + \lambda_k^3)}{\sqrt{n}} e^{-\lambda_k^2/2} + \frac{1}{n^{\frac{p-2}{2}}} \right]$$

where c_5 is a constant depending on p, c_1, c_2 . Thus

$$\log P(\lambda_k) = k \log [\Phi(\lambda_k) - \Phi(-\lambda_k)] + Z_k,$$

where

$$|Z_k| \leq \sum_{r=1}^k \left| \log \left[1 + \frac{e_{rk}}{\Phi(\lambda_k) - \Phi(-\lambda_k)} \right] \right|.$$

Since for sufficiently large k

$$\frac{|e_{rk}|}{\Phi(\lambda_k) - \Phi(-\lambda_k)} < \frac{1}{2}$$

and $\log(1+x) = x + vx^2, |v| < 1$ for $|x| < \frac{1}{2}$,

$$|Z_k| \leq c_6 k \left[\frac{1 + \lambda_k^3}{\sqrt{n}} e^{-\lambda_k^2/2} + \frac{1}{n^{\frac{p-2}{2}}} \right]. \quad \dots (3.1)$$

Substituting for λ_k in (3.1) we have

$$|Z_k| \leq c_6 \left[\frac{1 + \lambda_k^3}{\sqrt{n}} \cdot \sqrt{\log k} \cdot e^{-\theta} + \frac{k}{n^{\frac{p-2}{2}}} \right]. \quad \dots (3.2)$$

Since $n \geq (k \log k)^{2/p-2}$ it is easy to see from (3.2) that $|Z_k| \rightarrow 0$ as $k \rightarrow \infty$. This completes the proof.

The statistic τ_{nk} defined above can be used only when $\beta_2(x)$ is known. The natural way of modifying this in the case of unknown variance function is to replace β_{2r} by $\bar{y}_r^2 - \bar{y}_r^2$ where $\bar{y}_r^2$ and $\bar{y}_r$ are as in (2.1). Since this leads to certain complications in evaluating the limiting distribution we replace it by

$$s_r^2 = \frac{1}{n} \sum_{s=1}^n [y_{rs} - \mu(x_{rs})]^2 \text{ (where } x_{rs} \text{ is as in (2.1)) and write}$$

$$t_{nk} = \sup_{r=1, 2, \dots, k} \sqrt{n} \cdot \frac{|\bar{y}_r - \bar{\mu}_r|}{s_r}$$

To show that this replacement does not change τ_{nk} effectively in large samples we prove the following lemma.

Lemma 4: If X has a continuous distribution function and Y is a bounded random variable, $\beta_2(x) > c_1 > 0$ and if $n \geq (\log k)^{3+\delta}$, $\delta > 0$ then under the null hypothesis H_0

$$\lim_{k \rightarrow \infty} \log k \cdot \sup_r \left| \sqrt{\frac{s_r^2}{\beta_{2r}}} - 1 \right| = 0,$$

with probability one.

Proof: We have for any $\epsilon > 0$,

$$\begin{aligned} P \left[\sup_r \left| \sqrt{\frac{s_r^2}{\beta_{2r}}} - 1 \right| > \frac{\epsilon}{\log k} \mid H_0 \right] &\leq \sum_{r=1}^k P \left[\left| \frac{s_r^2}{\beta_{2r}} - 1 \right| > \frac{\epsilon}{\log k} \right] \\ &\leq \sum_{r=1}^k P \left[|s_r^2 - \beta_{2r}| > \frac{\epsilon c_1}{\log k} \right] \\ &\leq \sum_{r=1}^k P \left[\left| \bar{y}_r^2 - \bar{y}_r^2 \right| > \frac{\epsilon c_1}{2 \log k} \right] + \sum_{r=1}^k P \left[\left| \sum_{s=1}^n \varphi_{rs} y_{rs} - \varphi_{rs}^2 \right| > \frac{n \epsilon c_1}{4 \log k} \right]. \end{aligned}$$

Then from Theorem A2 we get

$$\sum_{r=1}^k P \left[\left| \bar{y}_r^2 - \bar{\xi}_r \right| > \frac{\epsilon c_1}{2 \log k} \mid x_1, \dots, x_{kn} \right] \leq 2k \exp \left[\frac{-b_1 n}{(\log k)^2} \right]$$

$$\text{and} \quad \sum_{r=1}^k P \left[\left| \sum_{s=1}^n (\varphi_{rs} y_{rs} - \varphi_{rs}^2) \right| > \frac{n \epsilon c_1}{4 \log k} \right] \leq 2k \exp \left[-\frac{b_2 n}{(\log k)^2} \right]$$

where b_1 and b_2 are constants depending on c_1, ϵ and the upper bound of $|Y|$. Thus whenever $n \geq (\log k)^{3+\delta}$ the series

$$\sum_{k=1}^{\infty} P \left[\sup_r \left| \sqrt{\frac{s_r^2}{\beta_{2r}}} - 1 \right| > \frac{\epsilon}{\log k} \right]$$

converges. An application of Borel-Cantelli lemma completes the proof of the lemma.

Remark: It is possible to show that if $n \geq k (\log k)^{2+\delta}$, $\delta > 0$ then $\log k \sup_r \left| \sqrt{\frac{s_r^2}{\beta_{2r}}} - 1 \right|$ converges to zero in probability even when Y is not a bounded random variable.

The following theorem enables us to use the statistic t_{nk} for testing a specified regression function when the conditional variance function is not known.

Theorem 3: *If X is a random variable with a continuous distribution in $[0, 1]$, Y is a bounded random variable and $\beta_2(x) > c_1 > 0$ then for $n \geq (k \log k)^{2/p-2}$ for some $p \geq 3$ we have*

$$\lim_{k \rightarrow \infty} P[t_{nk} < \lambda_k \mid H_0] = \exp \left[-\frac{1}{\sqrt{\pi}} e^{-\theta} \right]$$

where

$$\lambda_k = \sqrt{2(\theta + \log k - \frac{1}{2} \log \log k)}.$$

Proof: It is easily seen that

$$P[\tau_{nk} < \lambda_k(1 + V_{nk}) \mid H_0] \leq P[t_{nk} < \lambda_k \mid H_0] \leq P[\tau_{nk} < \lambda_k(1 + U_{nk}) \mid H_0] \quad \dots (3.3)$$

where τ_{nk} is as in Theorem 2, $U_{nk} = \sup_r \left(\sqrt{\frac{s_r^2}{\beta_{2r}}} - 1 \right)$ and $V_{nk} = \inf_r \left(\sqrt{\frac{s_r^2}{\beta_{2r}}} - 1 \right)$.

$$\text{Further, } P[\tau_{nk} < \lambda_k(1 + V_{nk})] = P \left[\frac{\tau_{nk}^2 - 2 \log k + \log \log k}{2} \cdot Z_{nk} + v_{nk} \leq \theta \right]$$

where Z_{nk} converges to unity in probability and v_{nk} converges to zero in probability because of Lemma 4. This fact together with an application of Theorem 2 leads to the result

$$\lim_{k \rightarrow \infty} P[\tau_{nk} < \lambda_k(1 + V_{nk})] = \exp \left[-\frac{1}{\sqrt{\pi}} e^{-\theta} \right]. \quad \dots (3.4)$$

A similar procedure leads to the result

$$\lim_{k \rightarrow \infty} P[\tau_{nk} < \lambda_k (1 + U_{nk})] = \exp \left[-\frac{1}{\sqrt{\pi}} e^{-\theta} \right]. \quad \dots (3.5)$$

(3.3), (3.4) ad (3.5) complete the proof.

Finally we shall state a theorem concerning the order of change in the level of significance if we apply the test 'reject H_0 if and only if $\tau_{nk} > \lambda_k$ ' where $[\Phi(\lambda_k) - \Phi(-\lambda_k)]^k = 1 - \alpha$ instead of the test 'reject H_0 if and only if $\tau_{nk} < \tau_{nk}(\alpha)$ with $P[\tau_{nk} \leq \tau_{nk}(\alpha) | H_0] = 1 - \alpha$.

Theorem 4: Let λ_k satisfy $[\Phi(\lambda_k) - \Phi(-\lambda_k)]^k = 1 - \alpha$ and $\alpha < 1 - \left[1 - \sqrt{\frac{2}{e\pi}} \right]^k$.

Then under the conditions of Theorem 2 we have

$$\left| P[\tau_{nk} \leq \lambda_k | H_0] - (1 - \alpha) \right| \leq c \cdot \left[\frac{\left(\log \frac{k}{\alpha} \right)^2}{\sqrt{n}} \cdot \alpha + \frac{k}{n^{\frac{p-2}{2}}} \right]$$

where c is a constant depending on p , c_1 and c_2 .

4. REMARKS

In this paper we have not given any consideration to the power of the τ_{nk} and t_{nk} tests proposed in Section 3. It would be very interesting if lower bounds for the power of these tests could be given in terms of the supremum distance between the regression functions under the null hypothesis and the alternative hypothesis. Though this has not been done, we can at least easily verify that if the regression functions under null hypothesis and the alternative hypothesis are both continuous, the τ_{nk} and t_{nk} tests are consistent, i.e., for any given level of significance $0 < \alpha < 1$, the probability of rejecting H_0 tends to unity as $k \rightarrow \infty$.

For the validity of the above theorems we have imposed certain conditions on the regression function φ and the variance function β_2 separately. In some cases (e.g. Binomial on Poisson distribution) β_2 is a function of φ . This however does not affect our analysis in any way so long as the conditions on φ and β_2 remain valid separately.

For simplicity in proof we assumed that $\beta_2(x)$ and $\beta_m(x)$ are bounded away from zero and bounded above respectively. However, it is easy to show that theorems 1, 2 and 4 are valid under the weaker assumption that $E[\beta_2(x)]^{-3(p-2)/2}$ and $E[\beta_3(x)]^{6(p-2)/p}$ are finite.

* $\Phi(x)$ is the usual normal distribution function.

Appendix

In this section we shall state and sketch the proof of some of the results which were utilised in our paper.

Let $X_1, X_2, X_3 \dots$ be a sequence of independent random variables with $EX_n = 0$ and $EX_n^2 = \sigma_n^2$ and β_{vn} the v -th absolute moment of X_n . We shall suppose that the moments of order $k > 3$ exist. We write

$$B_{vn} = \frac{1}{n}(\beta_{v1} + \dots + \beta_{vn}), \quad \rho_{vn} = \frac{B_{vn}}{B_{2n}^{v/2}}, \quad T_{kn} = \frac{\sqrt{n}}{4\rho_{kn}^{3/k}} \quad \dots \quad (A1)$$

$$\text{Then} \quad 1 \leq \rho_{vn} \leq \rho_{kn}^{v/k}, \quad v = 2, 3, \dots, k. \quad \dots \quad (A2)$$

Let $f_n(t)$ and $F_n(x)$ be the characteristic function and distribution function of $\frac{X_1 + \dots + X_n}{\sqrt{nB_n}}$. We now state two lemmas one of which is due to Cramér (1937) and the other due to Esseen and Berry.

Lemma A1 : (Cramér). For $|t| \leq T_{kn}^{1/3}$ we have

$$\left| f_n(t) - e^{-t^2/2} - \sum_{v=1}^{k-3} \frac{P_{vn}(it)}{n^{v/2}} e^{-t^2/2} \right| \leq \frac{c_k}{T_{kn}^{k-2}} [|t|^k + |t|^{3(k-2)}] \cdot e^{-t^2/2}$$

where

$$P_{vn}(it) = \sum_{j=1}^v c_{jvn} (it)^{v+2j},$$

$$c_{jv} = c'_k \cdot \rho_{kn}^{\frac{v+2j}{k}} \text{ and } c_k \text{ and } c'_k \text{ are constants depending only on } k, c_k \text{ being positive.}$$

Lemma A2 : (Berry-Esseen). For $|t| \leq T_{3n}$

$$\left| f_n(t) - e^{-t^2/2} \right| \leq \frac{4}{T_{3n}} \cdot |t|^3 \cdot e^{-t^2/2}.$$

Let

$$g(t) = e^{-t^2/2} \left[1 + \sum_{v=1}^{k-3} \frac{P_{vn}(it)}{n^{v/2}} \right]. \quad \dots \quad (A3)$$

Hereafter we shall denote by θ_k any positive constant which depends only on k . Then by using (A1), (A2) and Lemma A1 we have

$$\begin{aligned} \left| \sum_{v=1}^{k-3} \frac{P_{vn}(it)}{n^{v/2}} \right| &\leq \theta_k \sum_{v=1}^{k-3} \sum_{j=1}^v \frac{c_{jvn}}{n^{v/2}} |t|^{v+2j} \leq \theta_k \cdot \sum_{v=1}^{k-3} \sum_{j=1}^v \frac{\rho_{kn}^{\frac{v+2j}{k}}}{n^{v/2}} \cdot |t|^{v+2j} \\ &\leq \theta_k \cdot \sum_{v=1}^{k-3} \sum_{j=1}^v \frac{\rho_{kn}^{\frac{3v}{k}}}{(\sqrt{n})^v} |t|^{v+2j} \leq \theta_k \sum_{v=1}^{k-3} \sum_{j=1}^v \frac{|t|^{v+2j}}{T_{kn}^v}. \quad \dots \quad (A4) \end{aligned}$$

From (A3) and (A4) it is seen that the inequality of Lemma A2 can be rewritten as

$$\left| f_n(t) - g_n(t) \right| \leq \frac{4}{T_{3n}} |t|^3 \cdot e^{-t^2/3} + \theta_k \cdot \sum_{v=1}^{k-3} \sum_{j=1}^v \frac{|t|^{\nu+2j}}{T_{kn}^\nu} e^{-t^2/2} \quad \dots \quad (A5)$$

Lemma A3 :

$$\int_{-T_{kn}}^{+T_{kn}} \left| \frac{f_n(t) - g_n(t)}{t} \right| dt \leq \frac{\theta_k}{T_{kn}^{k-2}}.$$

Case 1 : Let $T_{kn} \leq 1$, then $T_{kn} \leq T_{kn}^{1/3}$. Hence applying Lemma A1 we get

$$\int_{-T_{kn}}^{+T_{kn}} \left| \frac{f_n(t) - g_n(t)}{t} \right| dt \leq \frac{\theta_k}{T_{kn}^{k-2}}.$$

Case 2 : Let $T_{kn} > 1$. From (A2) it is easy to see that $1 < T_{kn} \leq T_{3n}$:

Thus

$$\int_{-T_{kn}}^{T_{kn}} \left| \frac{f_n(t) - g_n(t)}{t} \right| dt = 2 \int_{T_{kn}^{1/3}}^{T_{kn}} \left| \frac{f_n(t) - g_n(t)}{t} \right| dt + \int_{-T_{kn}^{1/3}}^{T_{kn}^{1/3}} \left| \frac{f_n(t) - g_n(t)}{t} \right| dt.$$

$$= 2I_1 + I_2 \text{ say.}$$

In the region of integration of I_1 Lemma A2 is applicable and for I_2 Lemma A1 is applicable. Thus from (A5) we have

$$\int_{-T_{kn}}^{T_{kn}} \left| \frac{f_n(t) - g_n(t)}{t} \right| dt \leq \frac{\theta_k}{T_{kn}^{k-2}} + \frac{\theta_k}{T_{kn}} e^{-T_{kn}^{2/3}/4} \leq \theta_k / T_{kn}^{k-2}.$$

Let

$$Q(x) \geq 0, 0 \leq q(t) \leq 1, \int_{-\infty}^{+\infty} Q(x) dx = 1.$$

$$q(t) = \int_{-\infty}^{+\infty} e^{itx} Q(x) dx, q(t) \leq \frac{c}{|t|^{3(k-2)}}$$

$$\int_{-\infty}^{+\infty} |t|^{3k-7} q(t) dt < \infty. \quad \dots \quad (A6)$$

Lemma A4 : If $q(t)$ satisfies conditions (A6) then for $|x| > 1$

$$P(T_{kn}) = \int_{-T_{kn}}^{+T_{kn}} \left| \frac{f_n(t) - g_n(t)}{t} \right| q(xt) dt < \frac{\theta}{1 + |x|^k} \cdot \frac{1}{T_{kn}^{k-2}}$$

SOME LIMIT THEOREMS IN REGRESSION THEORY

Proof: Case 1: Let $T_{kn} \leq 1$, then $T_{kn} \leq T_{kn}^{1/3}$. Applying Lemma A1.

$$P(T_{kn}) \leq \frac{\theta_k}{T_{kn}^{k-2}} \int_{-\infty}^{+\infty} [|t|^k + |t|^{3(k-2)}] e^{-t^2/2} \cdot \frac{q(xt)dt}{|t|}$$

$$\leq \frac{\theta_k}{T_{kn}^{k-2}} \cdot \frac{1}{|x|^k} \cdot \int_{-\infty}^{+\infty} [|t|^{k-1} + |t|^{3k-7}] e^{-t^2/2} q(t)dt \leq \frac{\theta_k}{(1+|x|^k)} \cdot \frac{1}{T_{kn}^{k-2}}.$$

Case 2: Let $T_{kn} > 1$. Then as before $1 < T_{kn} \leq T_{3n}$.

$$P(T_{kn}) = 2 \int_{T_{kn}^{1/3}}^{T_{kn}} + \int_{-T_{kn}^{1/3}}^{-T_{kn}} \left| \frac{f_n(t) - g_n(t)}{t} \right| q(xt)dt = 2I_1 + I_2 \text{ say.}$$

As before by applying Lemma A1 for I_2 and Lemma A2 for I_1 and proceeding as in Case 1 we obtain

$$P(T_{kn}) \leq \frac{\theta_k}{1+|x|^k} \cdot \frac{1}{T_{kn}^{k-2}}.$$

Utilising Lemmas A3 and A4 and proceeding along the same lines as Esseen (1944) we can prove the following results.

Theorem A1: Let $X_1, X_2, \dots, X_n$ be a sequence of random variables with mean zero and finite absolute moments of order k ($k \geq 3$). Then for $x \geq \lambda_n$

$$\left| F_n(x) - \Phi(x) - \sum_{\nu=1}^{k-3} \frac{P_{\nu n}(-\Phi)}{n^{\nu/2}} \right| \leq \frac{\theta_k}{1+|x|^k} \cdot \frac{1}{T_{kn}^{k-2}} \text{ and for } x < \lambda_n$$

$$\left| F_n(x) - \Phi(x) - \sum_{\nu=1}^{k-3} \frac{P_{\nu n}(-\Phi)}{n^{\nu/2}} \right| \leq \theta'_k \left[\frac{(1+|x|^3)e^{-x^2/2}}{T_{kn}} + \frac{1}{T_{kn}^{k-2}} \right]$$

where $\lambda_n = 1 + \delta$ if $T_{kn} \leq 1$ and $\max [1 + \delta, \sqrt{2 \cdot (1 + \delta)(k-2) \log T_{kn}}]$ if $T_{kn} > 1$.

θ_k and θ'_k are constants which depends only on k and δ , $P_{\nu n}(-\Phi) = \sum_{j=1}^{\nu} (-1)^{\nu+2j} c_{j\nu n}$, $\Phi^{(\nu+2j)}(x)$ and $c_{j\nu n}$ are as in Lemma A 1.

Remark: From the above theorem it can be easily deduced that

$$\left| F_n(x) - \Phi(x) \right| \leq \frac{\theta_k}{1+|x|^k} \cdot \frac{1}{T_{kn}^{k-2}} \text{ for } x > \lambda_n \quad \dots \quad (\text{A } 7)$$

$$\left| F_n(x) - \Phi(x) \right| \leq \theta'_k \left[\frac{(1+|x|^3)e^{-x^2/2}}{T_{kn}} + \frac{1}{T_{kn}^{k-2}} \right] \text{ for } x \leq \lambda_n. \quad (\text{A } 8)$$

Let $\xi_1, \xi_2, \dots, \xi_n$ be n independently distributed random variables with $E(\xi_i) = 0$, $|\xi_i| \leq c$, $i = 1, 2, \dots, n$, $E(\xi_1 + \dots + \xi_n)^2 = \sigma^2$ and $\xi = \xi_1 + \dots + \xi_n$. Then we have the following theorem due to Prohorov (1959).

Theorem A2 : Under the conditions stated above, for $x > 0$ we have

$$P[\xi > x] \leq \exp \left[-\frac{x}{2c} \sinh^{-1} \frac{xc}{2\sigma^2} \right].$$

REFERENCES

- CRAMÉR, H. (1937) : Random variables and probability distributions. *Cambridge Tracts in Mathematics*, No. 36.
- ESSEEN, C. G. (1944) : Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian Law. *Acta Math.*, 77, 1.
- MAHALANOBIS, P. C. (1958) : Lectures in Japan : Fractile Graphical Analysis. Indian Statistical Institute.
- PROHOROV, YU. V. (1959) : An extremal problem in probability theory. *Teoriia Veroyatnostei i ee Primeneniia* 4, 211.

Paper received : May, 1960.

ACKNOWLEDGEMENT

Thanks are due to the following for their kind help in the editing of the papers published in this issue.

Shri S. John

Dr. C. R. Rao

Professor D. B. Lahiri

Shri R. Ranga Rao

Dr. Sujit Kumar Mitra

Shri J. Sethuraman

Editor,

Sankhyā : The Indian Journal of Statistics.

BIOMETRIKA

Volume 48, Parts 1 and 2

June 1961

CONTENTS

Studies in the history of probability and statistics XI. Daniel Bernoulli on maximum likelihood.....	Kendall, M. G.
The variance of Spearman's rho in normal samples.....	David, F. N. and Mallows, C. L.
Tests for rank correlation coefficients. II.....	Fieller, E. C. and Pearson, E. S.
Some methods of constructing exact tests.....	Durbin, J.
Preemptive priority queueing.....	Heathcote, C. R.
A two-sample sequential t -test.....	Hajnal, J.
Absolute and incomplete moments of the multivariate normal distribution.....	Nabeya, S.
Asymptotic expansions for the mean and variance of the serial correlation coefficient.....	White, John S.
Significance tests for paired-comparison experiments.....	Starks, T. H. and David, H. A.
Goodness-of-fit tests on a circle.....	Watson, G. S.
The use of orthogonal polynomials of the positive and negative binomial frequency functions in curve fitting by Aitken's method.....	Gonin, H. T.
The estimation of regression and error-scale parameters, when the joint distribution of the errors is of any continuous form and known apart from a scale parameter.....	Verhagen, A. M. W.
Latent vectors of random symmetric matrices.....	Mallows, C. L.
Expected values of normal order statistics.....	Harter, H. Leon
A distribution analogous to the Borel-Tanner.....	Haight, Frank A.
Occupancy probability distribution critical points.....	Nicholson, W. L.
Test of independence in intraclass 2×2 tables.....	Okamoto, Masashi and Ishii, Goro
Miscellanea	
Reviews	
Other Books Received	
Corrigenda	

The subscription, payable in advance, is 54/- (or \$8.00), per volume (including postage). Cheques should be made payable to *Biometrika*, crossed "*A/c Biometrika Trust*" and sent to the Secretary, Biometrika Office, Department of Statistics, University College, London, W. C. 1. All foreign cheques must be drawn on a Bank having a London agency.

Issued by THE BIOMETRIKA OFFICE, University College, London.

JOURNAL OF THE ROYAL STATISTICAL SOCIETY

Series B (Methodological)

40s each Part. Annual Subscription £3.2s (US \$9). Post free

CONTENTS OF VOL. 23, PART 1, 1961

Consistency in statistical inference and decision (With Discussion).....	Smith, C. A. B.
Delays on a two-lane road.....	Tanner, J. C.
A queueing model for road traffic flow (With Discussion on the two papers).....	Miller, A. J.
An unbiased estimator for powers of the arithmetic mean.....	Glasser, G. J.
A bulk service queueing problem with variable capacity.....	Jaiswal, N. K.
A simple method of trend construction.....	Leser, C. E. V.
The real stable characteristic functions and chaotic acceleration.....	Good, I. J.
Reply to Mr. Quenouille's comments about my paper on mixtures.....	Scheffé, H.
An asymptotic efficiency of Daniel's generalised correlation coefficients.....	Farlie, D. J.
The average run length of the cumulative sum chart when a V -mask is used.....	Kemp, K. W.
The time-dependent solution for an infinite dam with discrete additive inputs.....	Yeo, G. F.
Confidence limits for multivariate ratios.....	Bennett, B. M.
Estimation of the parameters of a linear functional relationship.....	Dorf, M. and Gurland, J.
A note on vacancies on a line.....	Downton, F.
Inaccuracy and inference.....	Kerridge, D. F.
A simple congestion system with incomplete service.....	Cox, D. R.
A test of homogeneity for ordered variances.....	Vincent, S. E.
The solution of queueing and inventory models by semi-Markov processes.....	Fabens, A. J.
A note on the renewal function when the mean renewal life time is infinite.....	Smith, W. L.
The moment generating function of the truncated multi-normal distribution.....	Tallis, G. M.

Obtainable through any bookseller or direct from the

ROYAL STATISTICAL SOCIETY

21 Bentinck Street, London, W. 1.

THE ANNALS OF MATHEMATICAL STATISTICS

(Founded by H. C. Carver)

The Official Journal of the Institute of Mathematical Statistics

CONTENTS OF VOLUME 32, NO 1, MARCH, 1961

Charles Jordan, 1871-1959.....	Takacs, I.
Statistical methods in Markov chains.....	Billingsley, Patrick
The frequency count of a Markov chain and the transition to a continuous time.....	Good, I. J.
On the asymptotic distribution of the "Psi-Squared" goodness of fit criteria for Markov chains and Markov sequences.....	Bhat, B. R.
Some properties of regular Markov chains.....	Bhat, B. R.
Some tests for categorical data.....	Bhaskar, V. P.
Tables for unbiased tests on the variance of a normal population.....	Pachares, James
Asymptotic efficiency of certain locally most powerful rank tests.....	Capon, Jack
The nonparametric ordering: (1001)→(0110).....	White, John S.
The non-central multivariate beta distribution.....	Kshirsagar, A. M.
A unified theory of estimation, I.....	Birnbaum, Allan
Admissible and minimax estimates of parameters in truncated spaces.....	Katz, Morris W.
The method of moments applied to a mixture of two exponential distributions.....	Rider, Paul R.
Snowball sampling.....	Goodman, Leo A.
Probability content of regions under spherical normal distributions, III: the bivariate normal integral.....	Ruben, Harold
Recurrent games and the Petersburg paradox.....	Robbins Herbert
Consistency and limit distributions of estimators of parameters in explosive stochastic difference equations.....	Rao, M. M.
First emptiness of two dams in parallel.....	Gani, J.
The transient behavior of a coincidence variate in telephone traffic.....	Finch, P. D.
First passage times of a generalized random walk.....	Kinney, John R.
Identifiability of mixtures.....	Teicher Henry
An asymptotic formula for the differences of the powers at zero.....	Good, I. J.
On a theorem of Rényi concerning mixing sequences of sets.....	Abbott, J. H. and Blum, J. R.
Theorems concerning Eisenhart's model II.....	Graybill, Franklin A. and Hultquist, Robert A.
Randomization and factorial experiments.....	Ehrenfeld, S. and Zacks, S.
Optimum designs in regression problems, II.....	Kiefer, J.
Non-equivalent comparisons of experiments and their use for experiments involving location parameters.....	Stone, M.
Distribution of the likelihood ratio for testing multivariate linear hypotheses.....	Katti, S. K.
A bound for the law of large numbers for discrete Markov processes.....	Katz, Melvin, Jr. and Thomasian, A. J.
The non-absolute convergence of Gil Pelaez' inversion integral.....	Wendel, J. G.

Subscription Rate: \$ 10.00 per year

Address orders for subscriptions and back numbers to the

Secretary, Institute of Mathematical Statistics,

Department of Statistics, University of

North Carolina, Chapel Hill, North Carolina.

ECONOMETRICA

Vol. 29, No. 1

January, 1961

CONTENTS

Economic change and rural-urban land conversions.....	Muth, Richard F.
Commodity group expenditure functions for the United Kingdom, 1948-1957.....	Leser, C. E. V.
Stability in voting.....	Dummett, Michael and Farquharson, Robin
Investment behaviour in the Japanese cotton spinning industry, 1916-1934.....	Ueno, Hiroya
The desirability of price instability under perfect competition.....	Ol, Walter Y.
A note on serial correlation bias in estimates of distributed lags.....	Griliches, Zvi
Oskar Anderson, 1887-1960.....	Fels, Eberhard
Georges Darmon, 1888-1960.....	Roy, Rene
Hans Peter's contribution to economics and econometrics.....	Haller, Heinz
Book Reviews	
Report of the Council, December, 1960	
Election of Fellows, 1960	
Fellows of the Econometric Society	
News and Notes	
Errata	

Subscription price for nonmembers : \$ 12.00 per year in the United States and Canada;

\$ 10.00 per year in all other Countries. Single copies \$ 3.00

Graduate Students recommended by a faculty member : \$ 4.00 per year.

Subscriptions should be sent to

**The Secretary
ECONOMETRIC SOCIETY**

Box 1264, Yale Station, New Haven, Connecticut, U.S.A.

SANKHYĀ

THE INDIAN JOURNAL OF STATISTICS

Edited by : P. C. MAHALANOBIS

SERIES A, VOL. 23

MAY 1961

PART 2

THE WEIGHTED MEAN OF TWO NORMAL SAMPLES WITH UNKNOWN VARIANCE RATIO

SIR RONALD A. FISHER, F.R.S.

Division of Mathematical Statistics, CSIRO, Adelaide

SUMMARY. The exact small-sample solution of the problem of the weighted mean, flows from an analysis parallel with that required for Behrens' problem. The algebraic forms are, however, more complex, involving one more parameter, for which it is convenient to choose Sukhatme's d . The logical basis of choice between weighted and unweighted means is considered.

1. ANALYTICAL PROCEDURE

Since the analytic demonstration of Behrens' test of significance for a difference between the means of two Normal samples has been set out explicitly (Fisher, 1935), it has been apparent that the problem of the distribution of the weighted mean of two such samples could be resolved, in the exact terms appropriate to small samples, by a similar analysis. In its mathematical generality, however, this problem is a more difficult one than the simple test of significance of the difference, for in addition to involving the two degrees of freedom, n_1 and n_2 , for the two samples, the modular angle

$$\tan \theta = s_1/s_2$$

of the ratio of the two standard deviations, and the level of significance required, it involved also as a fifth parameter, the measure of discrepancy of Behrens' test as defined by Sukhatme, namely

$$d = (\bar{x}_1 - \bar{x}_2) / \sqrt{s_1^2 + s_2^2}$$

and this to no unimportant an extent.

In a recent paper in *Sankhyā* (Fisher, 1961) I have shown how this test, like that of Behrens, can be exhibited as a verifiable assertion about frequencies, as must indeed always be the case when the Reference Set adverted to in any statements of probability has been explicitly specified.

Whereas Behrens' distribution is symmetrical, and so can be conveniently tabulated by giving the deviate, positive or negative, determining limits outside which d shall fall with a specified frequency, in random samples from the reference set, as is the case with Student's t , the distribution of the weighted mean is unsymmetrical in the general case. The analytical method of asymptotic expansion, which will be used in this paper, is however, just as applicable as it is in the case of Behrens' test (Fisher, 1941).

2. SPECIFICATION OF THE ANALYSIS

The ordinate of a Student's random variable t , for n degrees of freedom, may be expanded in Student polynomials, P_r in the form

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} \sum_0^{\infty} P_r n^{-r}, \quad \dots (1)$$

in which the first six polynomials are

$$\begin{array}{ll} P_0 = 1 & \\ P_1 = t^4 - 2t^2 - 1 & \div 4 \\ P_2 = 3t^8 - 28t^6 + 30t^4 + 12t^2 + 3 & \div 96 \\ P_3 = t^{12} - 22t^{10} + 113t^8 - 92t^6 - 33t^4 - 6t^2 + 15 & \div 384 \\ P_4 = 15t^{16} - 600t^{14} + 7100t^{12} - 26616t^{10} + 18330t^8 + 6360t^6 + 1980t^4 & \\ \quad - 1800t^2 - 945 & \div 92160 \\ P_5 = 3t^{20} - 190t^{18} + 4025t^{16} - 3397t^{14} + 103702t^{12} - 63444t^{10} - 21270t^8 & \\ \quad - 7800t^6 + 4455t^4 + 1890t^2 - 17955. & \div 368640 \end{array} \left. \vphantom{\begin{array}{l} P_0 \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \end{array}} \right\} \dots (2)$$

The simultaneous distribution of two Student variates is exhibited by the product of two expressions of the form (1).

The distribution of the weighted mean is determined from the Statistics observable in the two samples:

$$\left. \begin{array}{l} \bar{x}_1, s_1^2 = \frac{1}{n_1(n_1+1)} S(x_1 - \bar{x}_1)^2, \\ \bar{x}_2, s_2^2 = \frac{1}{n_2(n_2+1)} S(x_2 - \bar{x}_2)^2, \end{array} \right\} \dots (3)$$

by putting

$$\bar{x} = (s_2^2 \bar{x}_1 + s_1^2 \bar{x}_2) / (s_1^2 + s_2^2)$$

and

$$\frac{1}{S^2} = \frac{1}{s_1^2} + \frac{1}{s_2^2}, \quad \dots (4)$$

and obtaining finally

$$\mu_p = \bar{x} + S \sum_0^{\infty} \sum_0^{\infty} u_{rs} n_1^{-r} n_2^{-s}. \quad \dots (5)$$

WEIGHTED MEAN OF TWO SAMPLES WITH UNKNOWN VARIANCE RATIO

To obtain explicit expressions for u_{rs} , for any given probability we set

$$\mu - \bar{x}_1 = t_1 s_1$$

$$\mu - \bar{x}_2 = t_2 s_2, \quad s_1/s_2 = \tan \theta$$

whence
$$\mu - \bar{x} = \frac{s_1 s_2}{s_1^2 + s_2^2} (s_2 t_1 + s_1 t_2) = S(t_1 \cos \theta + t_2 \sin \theta). \quad \dots (6)$$

Let

$$t_1 \cos \theta + t_2 \sin \theta = \xi \quad \dots (7)$$

$$t_1 \sin \theta - t_2 \cos \theta = d$$

where d is Sukhatme's criterion for the significance of the observed difference between $\bar{x}_1$ and $\bar{x}_2$. Since d is known, the distribution of μ is required conditional upon a given value of d .

3. INTEGRATION

We should then substitute

$$\left. \begin{aligned} t_1 &= u \cos \theta + d \sin \theta \\ t_2 &= u \sin \theta - d \cos \theta \end{aligned} \right\} \quad \dots (8)$$

in Student's polynomials. Then

$$\int_{\xi}^{\infty} du \div \int_{-\infty}^{\infty} du$$

will give the probability of the deviation of the weighted mean (standardized by dividing by S) exceeding any chosen value ξ , the integrand being the product of two expressions in the form (1) for n_1 and n_2 degrees of freedom respectively.

From the way the variates have been transformed from t_1, t_2 to u, d the external factor

$$e^{-\frac{1}{2}(t_1^2 + t_2^2)} \quad \dots (9)$$

becomes
$$e^{-\frac{1}{2}u^2} \cdot e^{-\frac{1}{2}d^2} \quad \dots (10)$$

of which the latter factor is common both to numerator and denominator, and may, therefore, be omitted with the corresponding constant divisor $\sqrt{2\pi}$. The integral of

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} P_r(u \cos \theta + d \sin \theta) P_s(u \sin \theta - d \cos \theta) \quad \dots (11)$$

being that of the exponential multiplied by a polynomial in u , when taken from $-\infty$ to ∞ is a function of d and θ which may be designated by D_{rs} , so that the complete divisor is

$$D = \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} D_{rs} n_1^{-r} n_2^{-s}. \quad \dots (12)$$

We may at once note, writing c and s for the cosine and sine of θ ,

$$D_{00} = 1$$

$$\left. \begin{aligned} D_{10} &= \{4c^2s^2(d^2-1) + s^4(d^4-2d^2-1)\} \div 4 \\ D_{01} &= \{c^4(d^4-2d^2-1) + 4c^2s^2(d^2-1)\} \div 4 \end{aligned} \right\} \dots (13.1)$$

D_{20} , in tabular form

	1	d^2	d^4	d^6	d^8	
c^6s^2	-192	192	.	.	.	} $\div 96$
c^4s^4	48(+3	-18	+ 5	.	.)	
c^2s^6	8(+3	+27	-45	+ 7	.)	
s^8	+3	+12	+30	-28	+3	

... (13.2)

while D_{11} expressed similarly is

	1	d^2	d^4	d^6	d^8	
c^6s^2	4(-3	21	- 11	1	.)	} $\div 16$
c^4s^4	81	-252	122	-20	1	
c^2s^6	4(-3	21	- 11	1	.)	

... (13.3)

Expressions for D_{30} , D_{21} , D_{12} , D_{03} are easily found, but will only be needed if adjustment beyond the third are to be calculated.

The incomplete integral in the numerator involves terms of the form,

$$\int_{\xi}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} u^m du, \quad \dots (14)$$

which when m is odd is simply

$$z\{\xi^{m-1} + (m-1)\xi^{m-3} + \dots + (m-1)(m-3)\dots 2\}, \quad \dots (15)$$

but when m is even has in addition

$$q\{(m-1)(m-3) \dots 3\}, \quad \dots (16)$$

in which expressions z stands for

and q for	$\left. \begin{aligned} &\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\xi^2} \\ &\int_{\xi}^{\infty} z du \end{aligned} \right\}$	... (17)
-------------	---	----------

WEIGHTED MEAN OF TWO SAMPLES WITH UNKNOWN VARIANCE RATIO

4. DIVISION

We may, therefore, express the incomplete integral of the numerator as

$$z \sum_r \sum_s I_{rs} n_1^{-r} n_2^{-s} + q \sum_r \sum_s D_{rs} n_1^{-r} n_2^{-s}, \quad \dots (18)$$

and if dividing by $D = 1 + \frac{1}{n_1} D_{10} + \frac{1}{n_2} D_{01} + \dots$... (19)

there remains the quotient $q + z \sum_r \sum_s q_{rs} n_1^{-r} n_2^{-s},$... (20)

then the first operation (division) takes the form

$$\left. \begin{aligned} q_{10} &= I_{10}, & q_{01} &= I_{01} \\ q_{20} &= I_{20} - q_{10} D_{10}, \\ q_{11} &= I_{11} - q_{10} D_{01} - q_{01} D_{10}, \\ q_{30} &= I_{30} - q_{20} D_{10} - q_{10} D_{20}, \\ q_{21} &= I_{21} - q_{11} D_{10} - q_{20} D_{01} - q_{10} D_{11} - q_{01} D_{20}. \end{aligned} \right\} \dots (21)$$

and so on, adjusting the direct integral I at each stage by subtracting all products of the same weight (r, s) of D and q .

5. ADJUSTMENT OF DEVIATION

The expression

$$z Q = z \sum_r \sum_s q_{rs} n_1^{-r} n_2^{-s} \quad \dots (22)$$

gives the excess probability beyond the deviate ξ defined in terms of q .

If the probability evaluated is exactly

$$q(x) = q(\xi - F)$$

where F is some function of ξ , then

$$\xi = x + F$$

and

$$z Q = q(\xi - F) - q(\xi)$$

or

$$z \left\{ F + \frac{1}{2} \xi F^2 + \frac{1}{6} (\xi^2 - 1) F^3 + \frac{1}{24} (\xi^3 - 3\xi) F^4 + \dots \right\} \quad \dots (23)$$

using the expansion in Hermite polynomials. Hence F may be expressed in terms of ξ as

$$F = Q - \frac{1}{2}\xi Q^2 + \frac{1}{6}(2\xi^2 + 1)Q^3 - \frac{1}{24}(6\xi^3 + 7\xi)Q^4 + \dots \quad (24)$$

from which may be derived

$$\begin{aligned} f_{10} &= q_{10}, & f_{01} &= q_{01}, \\ f_{20} &= q_{20} - \frac{1}{2}\xi q_{10}^2, \\ f_{11} &= q_{11} - \xi q_{10} q_{01}, \\ f_{30} &= q_{30} - \xi q_{10} q_{20} + \frac{1}{6}(2\xi^2 + 1)q_{10}^3, \\ f_{21} &= q_{21} - \xi(q_{10} q_{11} + q_{01} q_{20}) + \frac{1}{2}(2\xi^2 + 1)q_{10}^2 q_{01} \end{aligned} \quad \dots \quad (25)$$

$$\text{and} \quad x + \sum_r \sum_s f_{rs} n_1^{-r} n_2^{-s} \quad \dots \quad (26)$$

is the deviate cutting off exactly the assigned probability

$$\int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du,$$

but in which the polynomials in F are expressed in terms of ξ .

Finally to express the successive adjustments in terms wholly of the normal deviate, known in advance, denoted by x , we put x for ξ and calculate

$$U = F + \frac{1}{2} \frac{d}{dx} F^2 + \frac{1}{6} \frac{d^2}{dx^2} F^3 + \dots \quad \dots \quad (27)$$

with, in detail

$$\begin{aligned} u_{10} &= f_{10}, & u_{01} &= f_{01}, \\ u_{20} &= f_{20} + \frac{1}{2} \frac{d}{dx} f_{10}^2, \end{aligned} \quad \dots \quad (28)$$

$$u_{11} = f_{11} + \frac{d}{dx} (f_{10} f_{01})$$

and so on.

The three first adjustments calculated in this way are set out in the accompanying Table.

WEIGHTED MEAN OF TWO SAMPLES WITH UNKNOWN VARIANCE RATIO

TABLE OF THE FIRST THREE ADJUSTMENTS

I

The adjustments are successively of the 4-th, 8-th and 12-th degrees in $\cos \theta$ and $\sin \theta$, and of the 3-rd, 5-th and 7-th degrees in x and d

$$u_{10}/n_1 = \{(x^3+x)c^4 + 4d(x^2+1)c^3s + 2(3d^2-1)xc^2s^2 + 4d(d^2-1)cs^3\} \div 4n_1$$

$$u_{01}/n_2 = \{-4d(d^2-1)c^3s + 2(3d^2-1)xc^2s^2 - 4d(x^2+1)cs^3 + (x^3+x)s^4\} \div 4n_2$$

where, as always, u_{sr} may be obtained from u_{rs} , by interchanging c and s and reversing the sign of d , or by writing c for s , $(-s)$ for c as if $(\pi/2+\theta)$ were written for θ .

II

For higher adjustments it is convenient to separate the odd and even powers of d . The terms to be divided by $96n_1^2$ are

even					odd				
		x^5	x^3	x			x^4	x^2	1
c^8	+ 1	5	16	3					
c^6s^2	- 4	4	23	51	c^7s	+ 48d	1	3	2
	+ 12d ²	.	15	31	c^5s^3	- 96d	1	5	8
c^4s^4	+ 12	.	2	9		+ 32d ³	.	11	13
	- 24d ²	.	10	39	c^3s^5	+ 96d	.	1	3
	+ 372d ⁴	.	.	1		- 64d ³	.	5	13
c^2s^6	+ 144d ²	.	.	1		+ 192d ⁵	.	.	1
	- 240d ⁴	.	.	1	cs^7	+ 96d ³	.	.	1
						- 96d ⁵	.	.	1

with conjugate terms divided by $96n_2^2$, while with divisor $16n_1n_2$ they are

even					odd				
		x^5	x^3	x			x^4	x^2	1
c^4s^4	+ 1	7	24	69					
	- 40d ²	.	2	7	$(c^5s^3 - c^3s^5)$	- 8d	3	13	22
	+ 108d ⁴	.	.	1		+ 16d ³	.	6	13
						- 48d ⁵	.	.	1
$(c^6s^2 + c^2s^6)$	- 2	.	5	9					
	+ 6d ²	.	5	17	$(c^7s - cs^7)$	+ 16d	.	1	1
	- 48d ⁴	.	.	1		- 16d ³	.	1	1

III

The third adjustment has four parts in two complementary pairs for u_{30} and u_{21} respectively divided by $384n_1^3$

even							odd						
			x^7	x^5	x^3	x			x^6	x^4	x^2	1	
c^{12}	+	1	3	19	17	-15							
							$c^{11}s$	+	$26d$	4	23	39	24
$c^{10}s^2$	-	2	24	181	632	915							
	+	$6d^2$	.	77	368	475	c^9s^3	-	$64d$	8	55	165	198
								+	$192d^3$	.	9	33	28
c^8s^4	+	4	12	134	649	1695							
	-	$12d^2$	.	196	1146	2486							
	+	$4d^4$	.	.	965	2341	c^7s^5	+	$384d$	1	10	40	71
								-	$2048d^3$	.	3	14	19
								+	$768d^5$	.	.	7	9
c^6s^6	-	8	.	8	70	225							
	+	$24d^2$	.	56	480	1441							
	-	$40d^4$	.	.	250	839	c^5s^7	-	$384d$	.	1	7	15
	+	$4488d^6$	.	.	.	1		+	$128d^3$	.	21	146	286
								-	$10368d^5$	.	.	1	2
								+	$1920d^7$	.	.	.	1
c^4s^8	-	$96d^2$	.	.	10	51							
	+	$96d^4$	.	.	35	181	c^3s^9	-	$256d^3$	.	.	5	16
	-	$6624d^6$	.	.	.	1		+	$384d^5$	.	.	7	23
								-	$2304d^7$	.	.	.	1
c^2s^{10}	-	$960d^4$	.	.	.	1							
	+	$1344d^6$	.	.	.	1	cs^{11}	-	$384d^5$	.	.	.	1
								+	$384d^7$	.	.	.	1

Normal deviations for chosen probabilities in a single tail are given below for ready reference.

P	x	P	x
.5	0	.01	2.32635
.25	.67449	.005	2.57583
.10	1.28155	.0025	2.80703
.05	1.64485	.001	3.09023
.025	1.95996	.0005	3.29053

WEIGHTED MEAN OF TWO SAMPLES WITH UNKNOWN VARIANCE RATIO
and by $384n_1^2n_2$

even						odd					
		x^7	x^5	x^3	x		x^6	x^4	x^2	1	
$c^{10}s^2$	— 2	.	77	328	435	$c^{11}s$	+ 192d	.	1	3	2
	+ 6d ²	.	77	568	931		— 192d ³	.	1	3	2
	— 96d ⁴	.	.	15	31	$c^{9}s^3$	— 16d	28	251	819	1080
$c^{8}s^4$	+ 1	153	1177	5099	11355		+ 384d ³	.	10	49	57
	— 144d ²	.	28	183	427		— 384d ⁵	.	.	11	13
	+ 24d ⁴	.	.	515	1679	$c^{7}s^5$	+ 32d	52	377	1341	2148
	— 5952d ⁶	.	.	.	1		— 512d ³	.	27	136	209
$c^{6}s^6$	— 4	36	361	1580	4065		+ 384d ⁵	.	.	53	101
	+ 180d ²	.	35	224	607		— 3840d ⁷	.	.	.	1
	— 8d ⁴	.	.	3040	11099	$c^{5}s^7$	— 192d	4	39	145	246
	+ 17880d ⁶	.	.	.	1		+ 128d ³	.	93	487	818
$c^{4}s^8$	+ 12	.	14	103	225		— 384d ⁵	.	.	61	137
	— 120d ²	.	14	121	319		+ 7680d ⁷	.	.	.	1
	+ 4d ⁴	.	.	3085	11813	$c^{3}s^9$	+ 192d	.	3	17	24
	— 11808d ⁶	.	.	.	1		— 128d ³	.	15	103	161
$c^{2}s^{10}$	+ 144d ²	.	.	5	13		+ 6912d ⁵	.	.	1	2
	— 48d ⁴	.	.	25	93		— 2304d ⁷	.	.	.	1
	+ 1728d ⁶	.	.	.	1	c^{11}	+ 384d ³	.	.	1	1
							— 384d ⁵	.	.	1	1

6. DISCUSSION

It appears unmistakably that the quantity d , defined by Sukhatme for implementing Behrens' test of the significance of the difference between the two observed means, is a major factor in determining the precision with which the weighted mean, with appropriate adjustments, can be used for the estimation of the common mean, assumed by hypothesis, of the two populations sampled. The analysis implies that the result of applying Behrens' test is not such that the hypothesis of a common mean is abandoned; it does not imply that the difference is not formally significant at some of the standard levels.

The importance of d in the formulae here developed is that the experimenter is led to realize that, if the populations have a common mean, but at the same time there has by chance occurred a somewhat wide discrepancy between the means observed, then the precision with which this common mean can be estimated from the two samples available is much lower than it would have been had the two means come in close agreement. This is really to be expected for two reasons:

(a) If the two means are close together, the weights attached to each in making a common estimate make little difference to the estimate obtained, which will be in this case principally liable only to nearly equal errors in the same direction of the means of two independent samples. Whereas with a large discrepancy the value of the common estimate will be much affected by the relative importance attached to the two pieces of evidence, and this, with small samples, is not well determined.

(b) A large discrepancy between the two means, in relation to the variation within samples, is in itself evidence that the mean squares observed within samples are both too low, and that the precision will be overestimated if the discrepancy is ignored. This information is absent on the hypothesis that the true means have an unknown difference.

A numerical example will perhaps make the position clearer. If we wish to locate the point which the true common mean has a probability of one in forty of exceeding, we should take, on large sample theory the value

$$\bar{x} + (1.95996)S.$$

The coefficient of S is modified by the first three corrective terms as follows in the case $n_1 = 15$, $n_2 = 20$, $\cos^2 \theta = 2/3$, and a range of values of d , as shown below.

The successive values for the adjusted coefficient seem to show a satisfactory convergence for the values of (n_1, n_2) chosen. The coefficients after three adjustments seem usually to be correct to two places of decimals and in the central region to three. As in other cases, it is to be expected that convergence will be slower at higher levels of significance, but more rapid for larger samples, with which, however, the influence of d may still be very considerable.

WEIGHTED MEAN OF TWO SAMPLES WITH UNKNOWN VARIANCE RATIO

d	first adjustment	first adjusted coefficient	second adjustment	second adjusted coefficient	third adjustment	third corrected coefficient
3.0	.80853	2.76849	.15687	2.92536	.00226	2.92762
2.5	.62419	2.58415	.08344	2.66759	-.00968	2.65791
2.0	.45831	2.41827	.03568	2.45395	-.01012	2.44383
1.5	.31483	2.27479	.00861	2.28340	-.00623	2.27717
1.0	.19768	2.15764	-.00280	2.15484	-.00234	2.15250
0.5	.11078	2.07074	-.00435	2.06639	-.00028	2.06611
0.0	.05806	2.01802	-.00166	2.01636	.00006	2.01642
-0.5	.04345	2.00341	+.00013	2.00354	-.00018	2.00336
-1.0	.07089	2.03085	-.00323	2.02762	.00013	2.02775
-1.5	.14478	2.10424	-.01460	2.08964	.00095	2.09059
-2.0	.26758	2.22754	-.03506	2.19248	-.00007	2.19241
-2.5	.44470	2.40466	-.06345	2.34121	-.00880	2.33241
-3.0	.67957	2.63953	-.09618	2.54335	-.03556	2.50779

An experimenter who can obtain values of a physical constant by two different methods, the sources of error in which are unrelated, may wish to review his data from two distinct points of view :

(a) That he has full assurance that his theoretical formulation is correct, and that his experimental procedures are both free from systematic error. In this case any apparent discrepancy between the two mean values is ascribed confidently to random sampling errors only, and fiducial limits for the common mean at any chosen level of probability may be obtained by the formulae of the present paper.

(b) That he does not exclude the possibility that his two methods would lead, if indefinitely repeated, to two different average values; that he will admit this possibility without knowing to what such a discrepancy might be due, or being able to compensate or allow for it. In this case, a large value for Sukatme's d is not evidence of lower precision. Indeed, as the discrepancy is not known to be wholly due to errors of random sampling the *weighted* mean has no special merit.

A statistical formulation appropriate to this case is that if μ_1 and μ_2 are the two population means of the two methods, there is one relevant quantity of which estimation is possible, namely

$$(\mu_1 + \mu_2)/2,$$

which presents a simpler problem than the weighted mean, for its error curve is symmetrical, and is indeed supplied by Behrens' test of significance. The error of the estimate,

$$(\bar{x}_1 + \bar{x}_2)/2$$

being

$$(t_1 s_1 + t_2 s_2)/2,$$

which, since the t distribution is symmetrical, is just half the variate tabulated for Behrens' test, when multiplied by $\sqrt{s_1^2 + s_2^2}$.

ACKNOWLEDGEMENT

I am indebted to L. C. Payne for detailed checking, so far as the second adjustment from the unpublished portion of his Ph.D. thesis.

REFERENCES

- CORNISH, E. A. and FISHER, R. A. (1937): Moments and cumulants in the specification of distribution. *Revue de l'Institut International de Statistique*, **5**, 307-352.
- (1960): The percentile points of distributions having known cumulants. *Technometrics*, **2**, 209-226.
- FISHER, R. A. (1926): Applications of 'Student's' distribution. *Metron*, **5**, 90-112.
- (1935): The fiducial argument in statistical inference. *Ann. Eug.*, **6**, 391-398.
- (1941): The asymptotic approach to Behren's integral, with further tables for the d test of significance. *Ann. Eug.*, **II**, 141-172.
- (1961): Sampling the reference set. *Sankhyā*, Series A, **23**, 1-6.
- FISHER, R. A. and YATES, F. (1957): *Statistical Tables for Biological Agricultural and Medical Research*. (Fifth edition)
- PAYNE, L. C. (1957): Unpublished portion of Ph.D. thesis. Cambridge.

Paper received : December, 1960.

A NOTE ON MUTUALLY ORTHOGONAL LATIN SQUARES

By S. S. SHRIKHANDE

Banaras Hindu University, India

SUMMARY. It is proved that the existence of a set of $(n-3)$ Mutually Orthogonal Latin Squares (MOLS) of order n implies the existence of a complete set of $(n-1)$ such squares and hence the existence of a finite projective plane $PG(2, n)$.

1. INTRODUCTION

A Latin Square of order n , is an $n \times n$ matrix whose entries are from a set of n distinct symbols such that each symbol occurs exactly once in each row and column. Two Latin Squares $A = (a_{ij})$, $B = (b_{ij})$ of order n are mutually orthogonal if the n^2 ordered pairs (a_{ij}, b_{ij}) are all distinct. A set $A_1, A_2, \dots, A_m$ of Latin Squares of order n , is called orthogonal if A_i and A_j are orthogonal for all $i \neq j$. It is easy to show that $m \leq n-1$. An orthogonal set is said to be complete provided if $m = n-1$. Bose (1938) and Levi (1942) independently showed the equivalence of complete sets of such squares of order n to the finite projective planes $PG(2, n)$.

A Balanced Incomplete Block design (BIBD) is an arrangement of v elements into b sets of k ($< v$) distinct elements such that every pair of different elements occurs in exactly λ sets. Then it is easy to show that every element occurs in exactly r of these sets. The numbers v, b, r, k, λ are called the parameters of the design. It is known that the existence of a complete set of squares of order n is equivalent to the existence of a BIBD with parameters

$$v = n^2, \quad b = n^2 + n, \quad r = n + 1, \quad k = n, \quad \lambda = 1. \quad \dots \quad (1.1)$$

Bose and Nair (1939) generalised the notion of a BIBD to a Partially Balanced Incomplete Block design (PBIBD). These new designs with 2-associate classes have been classified by Bose and Shimamoto (1952) according to their association schemes. Let $v = n^2$ treatments be exhibited in a square array of order n . If a set of $i-2$ MOLS of order n exists, then the association scheme of a PBIBD for these n^2 treatments is said to be of type L_i , if any two treatments are 2-associates, if and only if they occur in the same row, or column or correspond to the same symbol of any one of the Latin Squares. Then from Bose and Shimamoto (1952) we have

$$n_1 = (n-1)(n-i+1), \quad n_2 = i(n-1) \quad \dots \quad (1.2)$$

$$p^1_{11} = (n-i)^2 + i - 2; \quad p^2_{11} = (n-i)(n-i+1)$$

2. MAIN RESULT

Theorem : Any set of $(n-3)$ MOLS of order n can be uniquely extended to a complete set of $n-1$ such squares, if $n \neq 4$.

Proof : Suppose $(n-3)$ MOLS of order n exist. Then we can get a PBIB with association scheme L_{n-1} by forming blocks of size n , corresponding to the rows, columns or the same symbols of each of the $(n-3)$ Latin Squares. The parameters of this design are given by

$$v = n^2, \quad b = n^2 - n, \quad r = n - 1, \quad k = n, \quad \lambda_1 = 0, \quad \lambda_2 = 1, \quad \dots \quad (2.1)$$

$$n_1 = 2(n-1), \quad n_2 = (n-1)^2; \quad p_{11}^1 = n-2, \quad p_{11}^2 = 2.$$

From Shrikhande (1959), it follows that the association scheme of the above design is of type L_2 , i.e. the n^2 treatments can be uniquely arranged into a square array of order n , such that any two treatments in the same row or same column are 1-associates, otherwise they are 2-associates. Forming $2n$ blocks corresponding to the rows and columns of this array we obviously get a BIBD with parameters

$$v = n^2, \quad b = n^2 + n, \quad r = n + 1, \quad k = n, \quad \lambda = 1$$

which in turn implies the existence of a complete set of squares of order n . It is obvious that if we apply the same correspondence which was used to obtain the blocks of (2.1), then the added n blocks corresponding to rows (columns) give a Latin Square, such that these Latin Squares are mutually orthogonal and also orthogonal to the previous set of $(n-3)$ such squares.

Using Bruck and Ryser's Theorem (1949), we have.

Corollary 1 : If $n \equiv 1$ or $2 \pmod{4}$ and the square-free part of n contains a prime $4t+3$, then there do not exist $(n-3)$ MOLS of order n .

Similarly in the notation of Silverman (1960), we have the following refinement of his Theorem 4.8.

Corollary 2 : If $n \equiv 1$ or $2 \pmod{4}$ and the square free part of n contains a prime $4t+3$, then $S^k(n)$ is not r -orthogonal for $k \geq n+r-3$

REFERENCES

- BOSE, R. C. (1938) : On the application of the properties of Galois Fields to the problem of construction of hyper-Graeco-Latin squares. *Sankhyā*, **3**, 323-338.
- BOSE, R. C. and NAIR, K. R. (1939) : Partially balanced incomplete block designs. *Sankhyā*, **4**, 337-372.
- BOSE, R. C. and SHIMAMOTO, T. (1952) : Classification and analysis of partially balanced incomplete block designs with two-associate classes. *J. Amer. Stat. Ass.* **47**, 151-184.
- BRUCK, R. H. and RYSER, H. J. (1949) : The non-existence of certain projective planes. *Canad. J. Math.* **1**, 88-93.
- LEVI, F. W. (1942) : *Finite Geometrical Systems*, University of Calcutta.
- SHRIKHANDE, S. S. (1959) : The uniqueness of the L_2 association scheme. *Ann. Math. Stat.*, **30**, 781-798.
- SILVERMAN, R. (1960) : A metrization for power-sets with applications to combinatorial analysis. *Canad. J. Math.*, **10**, 158-176.

Paper received : January, 1961.

A STUDY OF BIB DESIGNS WITH REPLICATIONS 11 TO 15

By C. RADHAKRISHNA RAO

Indian Statistical Institute

SUMMARY. The combinations of parameters for BIB designs with replications 11 to 15 have been listed and the actual solutions of the cyclic type have been given in a number of cases. It may be observed that Fisher and Yates (1953) give the solutions to BIB designs up to 10 replications only.

1. INTRODUCTION

Sir Ronald Fisher, during his visit to the Indian Statistical Institute in the winter of 1960-61, suggested to the author that a study may be undertaken of the BIB designs with replications 11 to 15. He also gave a list of possible combinations of the parameters v , b , k , r and λ for such designs, except those derivable by doubling a known design with half the number of replications. This list is given in Table 1 together with the reference numbers, which are assigned following the convention of Fisher and Yates (1953). The designs derivable from geometrical configurations, denoted by *o.s.* (orthogonal squares) and *o.c.* (orthogonal cubes) in Fisher and Yates (1953) and Fisher (1945), are given a more explicit representation, indicating the particular geometrical configurations from which the solutions are obtained. The last non-geometrical design listed in Fisher and Yates (1953) has the reference no. 31 and, therefore, the non-geometrical designs in the present list of Table 1 are numbered consecutively from 32, first with increasing values of v within a replication (r), and then with increasing values of r . The actual numbers are written only in cases where the solutions have been found.

It may be seen that in a number of cases denoted by (—), the solution is unknown. In some cases solutions are derived by the methods given by Bose (1939) and Rao (1945, 1946) in so far as they are applicable. Other solutions are obtained by trial and error. The impossibility of certain designs has been established by using the results of Bruck and Ryser (1949), Schützenberger (1949), Chowla and Ryser (1950), Shrikhande (1950), and Hall and Connor (1954). Since their results and the author's (Rao, 1944, 1945, 1946) on the cyclic representations of geometrical designs are published mostly in Journals on Mathematics the relevant theorems are quoted in the next section for ready reference.

Some of the solutions obtained by trial and error and listed in Table 3 can be derived as particular cases of general combinatorial problems. For instance, a *combinatorial assignment problem* of which no. 66 in Table 1 is a special case may be stated as follows. In an establishment, there are v officers, r departments and k types of jobs in each department. Every officer has to be assigned a job in each department such that, in any department there are equal numbers of officers in different jobs and that any two officers have common jobs in exactly λ departments. When $v = s^2$ and s is a prime power a solution exists with $\lambda = 1$. For $s = 6$, no solution is possible with $\lambda = 1$, but a solution can be found with $\lambda = 2$ (derivable from the BIBD design no. 66). It is of some interest to determine for any given number v the minimum λ for which the assignment problem is soluble. All the numbers can then be characterised by the associated minimum λ .

SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS : SERIES A

TABLE 1. COMBINATIONS OF THE PARAMETERS FOR BIB DESIGNS WITH REPLICATIONS 11 TO 15

v	b	k	r	λ	ref. no.	v	b	k	r	λ	ref. no.
12	44	3	11	2	32	15	35	6	14	5	62
12	33	4	11	3	33	22	77	4	14	2	63
12	22	6	11	5	34	22	44	7	14	4	— (d)
23	23	11	11	5	35	29	58	7	14	3	65
45	99	5	11	1	36	36	84	6	14	2	66 (d)
45	55	9	11	2	—	43	86	7	14	2	67 (d)
56	56	11	11	2	—	78	91	12	14	2	*(Th 2)
100	110	10	11	1	?	85	170	7	14	1	—
111	111	11	11	1	?	92	92	14	14	2	*(Th 1)
13	26	6	12	5	41	169	182	13	14	1	E (2, 13):1
19	57	4	12	2	42	183	183	14	14	1	P (2, 13):1
21	42	6	12	3	43	11	55	3	15	3	71
22	33	8	12	4	—	13	39	5	15	5	72
25	100	3	12	1	45	16	80	3	15	2	73
33	44	9	12	3	—	16	48	5	15	4	74
34	34	12	12	4	*(Th 1)	16	40	6	15	5	75
37	111	4	12	1	48	16	30	8	15	7	E (4, 2):3
45	45	12	12	3	—	21	35	9	15	6	76
55	66	10	12	2	*(Th 2)	26	65	6	15	3	77
61	122	6	12	1	—	28	42	10	15	5	—
67	67	12	12	2	*(Th 4)	31	155	3	15	1	P (4, 2):1
121	132	11	12	1	E (2, 11):1	31	93	5	15	2	79
133	133	12	12	1	P (2, 11):1	31	31	15	15	7	P (4, 2):3
27	117	3	13	1	E (3, 3):1	36	36	15	15	6	—
27	39	9	13	4	E (3, 3):2	43	43	15	15	5	—
27	27	13	13	6	53	46	69	10	15	3	—
40	130	4	13	1	P (3, 3):1	56	70	12	15	3	—
40	52	10	13	3	—	61	183	5	15	1	84
40	40	13	13	4	P (3, 3):2	71	71	15	15	3	—
53	53	13	13	3	*(Th 3)	76	190	6	15	1	—
66	143	6	13	1	56	91	195	7	15	1	87
66	78	11	13	2	—	91	105	13	15	2	*(Th 2)
79	79	13	13	2	—	106	106	15	15	2	*(Th 1)
144	156	12	13	1	?	136	204	10	15	1	—
157	157	13	13	1	?	196	210	14	15	1	*(Th 2)
15	42	5	14	4	61 (d)	211	211	15	15	1	*(Th 5)

—solution unknown, * solution does not exist, ? presumably non-existing, (d) double of insoluble type.

P and E have been used for PG, and EG, the Projective and Euclidean finite geometrical configurations whose construction is considered in Section 3.

A STUDY OF BIB DESIGNS WITH REPLICATIONS 11 TO 15

2. THEOREMS ON NON-EXISTENCE OF CERTAIN BIB DESIGNS

Theorem 1 : (Schützenberger, 1949; Chowla and Ryser, 1950; Shrikhande, 1950) : *For the existence of a symmetrical BIBD with parameters, v, b, k, r, λ , a necessary condition is that $(r-\lambda)$ is a perfect square, when v is even.*

Theorem 2 : (Hall and Connor, 1954; Shrikhande, 1960) : *The existence of the BIBD, $v-r, b-1, r-\lambda, r, \lambda$, implies that of the BIBD, $v = b, k = r, \lambda$, when $\lambda = 1$ or 2.*

For example, the designs 68 and 70 with the parameters

$$v = 78, b = 91, k = 12, r = 14, \lambda = 2$$

$$v = 92, b = 92, k = 14, r = 14, \lambda = 2$$

do not exist. By Theorem 1, the latter does not exist as $(k-\lambda) = 12$ is not a perfect square, and Theorem 2 excludes the possibility of the former. Similarly the designs 88 and 89 do not exist. No. 47 does not exist by Theorem 1, but this does not necessarily imply the non-existence of 44.

Theorem 3 : (Chowla and Ryser, 1950) : *If $v \equiv 1 \pmod{4}$ and there exists an odd prime p such that p divides the square free part of $k-\lambda$, and, moreover, if the Legendre symbol $(\lambda|p) = -1$, then the symmetric BIBD does not exist.*

Theorem 4 : (Chowla and Ryser, 1950) : *If $v \equiv 3 \pmod{4}$ and if there exists an odd prime p such that p divides the square free part of $k-\lambda$, and moreover, if $(-\lambda|p) = -1$, then the symmetric BIBD has no solution.*

For the design 52, ($v = 67 = b, r = 12 = k, \lambda = 2$), $v \equiv 3 \pmod{4}$, $k-\lambda = 10$ and $p = 5$ is an odd prime dividing the square free part of 10. We find $-\lambda = -2$ is a quadratic non-residue of 5, so that the Legendre symbol $(-2|5) = -1$. Hence, by Theorem 4, the symmetrical design 52 does not exist. By Theorem 2, 50 does not exist. By an application of Theorem 3, it can be shown in a similar way that 55 does not exist. This does not imply the non-existence of 54, as Theorem 2 is not applicable for $\lambda = 3$.

Theorem 5 : (Bruck and Ryser 1949) : *If $n \equiv 1$ or $2 \pmod{4}$ and p , a prime dividing the square free part of n , is of the form $4k+3$, then $(n-1)$ mutually orthogonal squares do not exist.*

Designs 91 and 92 are not possible for they imply the existence of 13 mutually orthogonal squares of order 14, which is impossible by Theorem 5.

3. CYCLIC DESIGNS DERIVABLE FROM GEOMETRICAL CONFIGURATIONS

A finite projective geometry of t dimensions with coordinates as elements of a Galois Field $GF(s)$ is represented by $PG(t, s)$ and the corresponding Euclidean geometry by $EG(t, s)$. Bose (1959) observed that by choosing the points as varieties and all d dimensional flats as blocks from either $PG(t, s)$ or $EG(t, s)$ one can generate a BIBD. Designs so obtained are represented by $PG(t, s) : d$ or $EG(t, s) : d$ under the reference number in Table 1. The parameters of such designs are :

	$PG(t, s) : d$	$EG(t, s) : d$
v	$(s^{t+1}-1)/(s-1)$	s^t
b	$\phi(t, s, d)$	$s^{t-d} \phi(t-1, s, d-1)$
k	$(s^{d+1}-1)/(s-1)$	s^d
r	$\phi(t-1, s, d-1)$	$\phi(t-1, s, d-1)$
λ	$\phi(t-2, s, d-2)$	$\phi(t-2, s, d-2)$

where
$$\phi(t, s, d) = \frac{(s^{t+1}-1) \dots (s^{t-d+1}-1)}{(s^{d+1}-1) \dots (s-1)}$$

Such geometries can be constructed for s equal to a prime or a prime power and, therefore, the designs in Table 1 with reference numbers, of the type $PG(t, s) : d$ and $EG(t, s) : d$ exist.

We shall demonstrate how cyclic solutions can be obtained in a simple manner for all designs derivable from finite geometrical configurations. The main results are contained in Rao (1944, 1945, 1946) where the theorems of Singer (1938) and Bose (1942) have been generalised and applied to the construction of cyclic solutions to BIB designs. The theorems quoted in this section are from Rao (1946), which is devoted to the construction of cyclic solutions to BIB designs derivable from cyclic representations of finite geometrical configurations.

Theorem 6 : Let $x^{t+1} - a_t x^t - \dots - a_0$ be a minimum function generating the elements of $GF(s^{t+1})$ and the sequence ξ_d , $[d = 0, 1, \dots, (s^{t+1}-s)/(s-1)]$, be derived from the recurrence relation

$$\xi_{d+1} = a_t \xi_d + a_{t-1} \xi_{d-1} + \dots + a_0 \xi_{d-t}$$

with the initial values $\xi_0 = \dots = \xi_{t-1} = 0$, $\xi_t = 1$. The set of integers d such that $\xi_d = 0$ constitutes a difference set mod $v = (s^{t+1}-1)/(s-1)$.

The number of d 's such that $\xi_d = 0$ is $k = (s^t-1)/(s-1)$ and among the $k(k-1)$ differences reduced mod v , each integer less than v occurs $\lambda = (s^{t-1}-1)/(s-1)$ times. The difference set provides a cyclic solution to the BIBD, $PG(t, s) : (t-1)$ with the parameters

$$v = \frac{s^{t+1}-1}{s-1} = b, \quad r = \frac{s^t-1}{s-1} = k, \quad \lambda = \frac{s^{t-1}-1}{s-1}$$

A STUDY OF BIB DESIGNS WITH REPLICATIONS 11 TO 15

The difference sets for $t = 2$ and $s = 3, 4, 5, 7, 8, 9$ are given in Fisher and Yates (1953) and also in Rao (1946). The solutions for the following are obtained using the minimum functions indicated.

ref. no.	parameters	minimum function
$PG(2,11) : 1$	$v = b = 133, r = k = 12, \lambda = 1$	$x^3 - 3x + 1$
$PG(2,13) : 1$	$v = b = 183, r = k = 14, \lambda = 1$	$x^3 + 2x + 2^*$
$PG(3, 3) : 2$	$v = b = 40, r = k = 13, \lambda = 4$	$x^4 - 2x^3 - 2x^2 - x - 1$
$PG(4, 2) : 3$	$v = b = 31, r = k = 15, \lambda = 7$	$x^5 - x^2 - 1$

We shall demonstrate the method for $PG(3,3) : 2$. The initial values are $\xi_0 = \xi_1 = \xi_2 = 0, \xi_3 = 1$ and the rest ξ_4 to ξ_{39} are obtained from the recurrence relation

$$\xi_{d+1} = 2\xi_d + 2\xi_{d-1} + \xi_{d-2} + \xi_{d-3}.$$

Thus $\xi_4 = 2, \xi_5 = 0, \xi_6 = 2, \dots$ and so on. The suffixes corresponding to zero values are

$$(0, 1, 2, 5, 12, 18, 22, 24, 26, 27, 29, 32, 33)$$

which constitute a difference set mod 40. On cyclic development we obtain the BIBD, $PG(3,3) : 2$, where the varieties are numbered 0 to 39. Similarly we find the cyclic solution to $PG(4, 2) : 3$

$$(0, 1, 2, 3, 5, 6, 8, 11, 12, 18, 19, 20, 23, 27, 29) \text{ mod } 31.$$

Two difference sets for this design obtained by alternative methods are

$$(1, 2, 4, 5, 7, 8, 9, 10, 14, 16, 18, 19, 20, 25, 28) \text{ mod } 31$$

due to Bose (1939) and

$$(1, 2, 3, 4, 6, 8, 12, 15, 16, 17, 23, 24, 27, 29, 30) \text{ mod } 31$$

due to Marshal Hall (1958). The last two are non-isomorphic and their relation to the first solution has not been investigated.

Theorem 7: Let $x^t - a_{t-1}x^{t-1} - \dots - a_0$ be a minimum function generating $GF(s^t)$, and the sequence $\xi_d, [d = 0, \dots, s^t - 1]$, be derived from the recurrence relation :

$$\xi_{d+1} = a_{t-1}\xi_d + \dots + a_0\xi_{d-t+1}$$

with the initial values $\xi_0 = \xi_1 = \dots = \xi_{t-2} = 0, \xi_{t-1} = 1$. The set of integers d such that $\xi_d = \alpha$ (fixed) $\neq 0$ provides a difference set such that among the differences mod $(s^t - 1)$, all integers less than $(s^t - 1)$ and not divisible by $\theta = (s^t - 1)/(s - 1)$ occur an equal number, (s^{t-2}) , of times and those divisible by θ , zero times.

*Minimum functions of certain orders are given in Charnichael (1937). Some algebraic methods of deriving minimum functions of the second order and, in a few cases, of the third order have been given by Bose, Chowla and Rao (1944, 1945a, 1945b). The third order minimum functions for $s = 13$ have not been reported anywhere in the literature. At the request of my colleague Dr. I. M. Chakravorty, Dr. Jack Alanen found some solutions with the help of a Burroughs-220 computer. I wish to thank Dr. Alanen for supplying a few minimum functions of the third order for $s = 13$, of which I am quoting one. A note by Dr. Alanen is appearing in this issue of *Sankhyā*.

Such a difference set provides a compact representation of the resolvable BIBD, $EG(t, s) : (t-1)$, with the parameters

$$v = s^t, b = (s^{t+1}-s)/(s-1), k = s^{t-1}, r = (s^t-1)/(s-1), \lambda = (s^{t-1}-1)/(s-1).$$

The s^t varieties are represented by the residues $0, 1, \dots, s^t-1$ and an element ∞ which is invariant under addition with the residues. Let $d_1, \dots, d_k$ be the difference set obtained as in Theorem 7. Consider the $(s-1)$ sets derived by adding $0, \theta, \dots, (s-2)\theta$ to each element of the initial set, i.e., in the cyclic development of $(d_1, \dots, d_k)$ we take the sets obtained after $\theta, 2\theta, \dots$ steps. This operation may be denoted by

$$[(d_1, \dots, d_k)S(\theta)] \bmod (s^t-1)$$

To this add another set consisting of the remaining $(k-1)$ residues of (s^t-1) and an invariant element ∞ . This operation is denoted by

$$[(d_1, \dots, d_k)S(\theta)+R] \bmod (s^t-1)$$

which gives one complete replication of the resolvable BIBD. The other replications are obtained by adding $1, 2, \dots, (\theta-1)$, to each member of the first replication. The procedure for developing the whole solution of the resolvable design may be represented as

$$PC(\theta)[(d_1, \dots, d_k)S(\theta)+R] \bmod (s^t-1)$$

where PC indicates a partial cycle, up to θ only.

As an example let us consider $EG(4, 2) : 3$ and the minimum function x^4-x^3-1 . We find

$$\begin{aligned} \xi_0 = \xi_1 = \xi_2 = \xi_7 = \xi_9 = \xi_{12} = \xi_{13} &= 0 \\ \xi_3 = \xi_4 = \xi_5 = \xi_6 = \xi_8 = \xi_{10} = \xi_{11} = \xi_{14} &= 1 \end{aligned}$$

and the set of d , such that $\xi_d = 1$,

$$3, 4, 5, 6, 8, 10, 11, 14$$

is a difference set mod 15 with the property stated in Theorem 7. One replication of the resolvable BIBD, $v = 16, b = 30, k = 8, r = 15, \lambda = 7$, is

$$[(3, 4, 5, 6, 8, 10, 11, 14)S(15)+R] \bmod 15$$

which on development yields one replication

$$(3, 4, 5, 6, 8, 10, 11, 14)$$

$$(\infty, 1, 2, 7, 9, 12, 13, 15)$$

and the other 14 replications are obtained by adding $1, 2, \dots, 14$ to the elements of the first replication and reducing to mod 15. Similarly solutions are obtained for the following and given in Table 2.

ref. no.	parameters of resolvable BIBD					minimum function
	v	b	k	r	λ	
$EG(2,11) : 1$	121	132	11	12	1	$x^2-4x+2, GF(11)$
$EG(3,2) : 2$	27	39	9	13	4	$x^3-x-2, GF(3)$
$EG(2,13) : 1$	169	182	13	14	1	$x^2-x-1, GF(13)$
$EG(4,2) : 3$	16	30	8	15	7	$x^4-x^3-1, GF(2)$

A STUDY OF BIB DESIGNS WITH REPLICATIONS 11 TO 15

Theorem 8: If s is a prime or a prime power and $\theta = (s^{t+1}-1)/(s^2-1)$ is not integral, it is possible to find $y = (s^t-1)/(s^2-1)$ sets

$$(d_{0j}, d_{1j}, \dots, d_{mj}), \quad j=1, \dots, y$$

such that all the differences mod $v = (s^{t+1}-1)/(s-1)$ contain integers less than v once and once only.

For the actual method of construction references may be made to Rao (1945, 1946). The solution obtained for $PG(4, 2) : 1$, with parameters, 31, 155, 3, 15, 1 is given in Table 2.

Theorem 9 : If s is a prime or prime power, there exist $\eta = (s^{t-1}-1)/(s-1)$ sets

$$(d_{1i}, \dots, d_{\eta i}), \quad i = 1, \dots, \eta$$

such that all the differences mod $v = (s^t-1)$ contain integers less than v and not divisible by $\theta = v/(s-1)$ once and those divisible by θ , zero times.

If we add to the difference set of Theorem 9 the set $(\infty, 0, \theta, \dots)$ with a partial cycle θ , we obtain a compact representation of the BIBD, $EG(t, s) : 1$, with the parameters

$$v = s^t, \quad b = s^{t-1}(s^t-1)/(s-1), \quad k = s, \quad r = (s^t-1)/(s-1), \quad \lambda = 1.$$

The solution for $EG(3,3) : 1$ is given in Table 2.

Theorems providing difference sets for the designs $PG(t, s) : d$ and $EG(t, s) : d$ for values of d other than those covered by Theorems 6, 7, 8, 9 are given in Rao (1945, 1946). They have not been quoted here as they do not provide solutions to any of the designs listed in Table 1.

TABLE 2. CYCLIC SOLUTIONS TO BIB DESIGNS DERIVABLE FROM FINITE GEOMETRICAL CONFIGURATIONS

ref. no.	solution
<i>Non-resolvable designs</i>	
$PG(2,11) : 1$	$(1, 2, 4, 13, 21, 35, 39, 82, 89, 95, 105, 110) \bmod 133$
$PG(2,13) : 1$	$(0, 1, 3, 24, 41, 52, 57, 66, 70, 96, 102, 149, 164, 176) \bmod 183$
$PG(3,3) : 1$	$(0, 1, 26, 32), (0, 7, 19, 36), (0, 3, 16, 38) \bmod 40, PC(10) (0, 10, 20, 30) \bmod 40$
$PG(3,3) : 2$	$(0, 1, 2, 5, 12, 18, 22, 24, 26, 27, 29, 32, 33) \bmod 40$
$PG(4,2) : 3$ (3 solutions)	(i) $(0, 1, 2, 3, 5, 6, 8, 11, 12, 18, 19, 20, 23, 27, 29) \bmod 31$
	(ii) $(1, 2, 4, 5, 7, 8, 9, 10, 14, 16, 18, 19, 20, 25, 28) \bmod 31$
	(iii) $(1, 2, 3, 4, 6, 8, 12, 15, 16, 17, 23, 24, 27, 29, 30) \bmod 31$
$PG(4,2) : 1$	$(0, 1, 18), (0, 2, 5), (0, 4, 10), (0, 8, 20), (0, 9, 16) \bmod 31$
$EG(3,3) : 1$	$(0, 1, 22), (0, 2, 8), (0, 3, 14), (0, 7, 17) \bmod 26, PC(13) (\infty, 0, 13) \bmod 26$
<i>Resolvable designs</i>	
$EG(2,11) : 1$	$PC(12)[(0, 9, 27, 29, 46, 50, 76, 104, 107, 114, 115)S(12) + R] \bmod 120$
$EG(3,3) : 2$	$PC(13)[(0, 1, 2, 8, 11, 18, 20, 22, 23) + R] \bmod 26$
$EG(2,13) : 1$	$PC(14)[(0, 9, 23, 35, 72, 92, 97, 110, 136, 151, 157, 158, 160)S(14) + R] \bmod 168$
$EG(4,2) : 3$	$PC(14)[(3, 4, 5, 6, 8, 10, 11, 14) + R] \bmod 15.$

4. CYCLIC SOLUTIONS TO NON-GEOMETRICAL DESIGNS

Bose (1939) discussed some methods of obtaining cyclic solutions and gave the actual solutions to several series of designs designated by $T_1, T_2, F_1, F_2, G_1, G_2, S_n$. Table 3 gives the cyclic solutions obtained by these methods, in so far as they are applicable in which cases the series to which each solution belongs has been indicated, and other solutions obtained by trial and error (indicated by the symbol —).

There are several types of cyclic solutions as may be seen from Table 3. In a simple cyclic solution such as 35, the varieties are represented by integers $0, 1, \dots, v-1$. From the initial set or sets, the whole design is generated by adding integers $1, 2, \dots, v-1$ and reducing to mod v . That is, an integer i in an initial set is changed to $i+1, \dots, v-1, 0, \dots, i-1$ in the derived sets. In some cases such as 32, the varieties are represented by integers and an invariant element designated as ∞ , (Fisher and Yates, 1953 use I for this purpose). This element remains the same in all the derived sets.

There are dicyclic solutions such as 43, where the varieties are represented by (x, y) , $x = 0, \dots, p-1$ and $y = 0, 1, \dots, q-1$. The cyclic development with respect to one of the coordinates is carried out first keeping the other fixed. From the sets so generated others are derived by a cyclic development of the other coordinate, fixed in the first operation. When the initial sets are given in such a way that no cyclic development is necessary with respect to one of the coordinates for obtaining the complete design a dash is indicated in the symbol, mod (p, q) as in 73 (ii). An additional complication in dicyclic solutions is the introduction of invariant elements as in 48 and 62. Some designs such as 36 have tricyclic solutions.

Notes on the designs in Table 3: 34 and 41 can also be obtained by block section from 35 and 53 respectively. It is not known whether from 76, one can build up the symmetrical design 80 with parameters 36, 36, 15, 15, 6, whose existence implies that of 76. Design 48, obtained by trial, is the most difficult one, and no known technique including the recent methods given by Bose and Shrikhande (1960) could yield a solution. A non-isomorphic solution to 32 is given by Skolem (1958).

Some of the designs in Table 3 are of the resolvable type. But the solutions given are not resolvable except in cases where the index (r) is shown with the reference number. Two solutions are listed for 32, of which only (i) is resolvable. In the case of a resolvable solution, either one complete replication, as in 32 and 33, or the method of generating one or more complete replications, as in 66, is shown within square brackets. All the replications are obtained by the indicated cyclic development of the initial replications. In the case of 66,

$$[(01, 06, 15, 12, 23, 24) \bmod (5, -) + R]$$

and

$$[(01, 06, 35, 32, 13, 14) \bmod (5, -) + R]$$

A STUDY OF BIB DESIGNS WITH REPLICATIONS 11 TO 15

TABLE 3. CYCLIC SOLUTIONS TO SOME BIB DESIGNS, NOT DERIVABLE FROM
FINITE GEOMETRICAL CONFIGURATIONS

ref. no.	v	b	k	r	λ	method	cyclic solution
32 ^r	12	44	3	11	2	—	(i) $[(0, 1, 3), (4, 5, 9), (2, 8, 6), (\infty, 7, 10)] \bmod 11$
		(two solutions)				E_1	(ii) $(0, 1, 3), (0, 1, 4), (0, 2, 6), (\infty, 0, 5) \bmod 11$
33 ^r	12	33	4	11	3	—	$[(0, 1, 3, 7), (2, 4, 9, 10), (\infty, 5, 6, 8)] \bmod 11$
34	12	22	6	11	5	—	$(0, 1, 3, 7, 8, 10), (\infty, 0, 5, 6, 8, 10) \bmod 11$
35	23	23	11	11	5	S_1	$(1, 2, 3, 4, 6, 8, 9, 12, 13, 16, 18) \bmod 23$
36	45	99	5	11	1	G_2	$(010, 020, 102, 202, 001), (210, 120, 222, 112, 001) \bmod (3, 3, 5)$
							$(000, 001, 002, 003, 004) \bmod (3, 3, -)$
41	13	26	6	12	5	—	$(0, 1, 3, 6, 7, 11), (0, 1, 2, 3, 7, 11) \bmod 13$
42	19	57	4	12	2	—	$(0, 1, 3, 12), (0, 1, 5, 13), (0, 4, 6, 9) \bmod 19$
43	21	42	6	12	3	—	$(00, 05, 14, 11, 22, 23), (00, 01, 03, 10, 11, 13) \bmod (3, 7)$
45	25	100	3	12	1	—	(i) $(0, 1, 3), (0, 4, 13), (0, 5, 11), (0, 7, 17) \bmod 25$
		(two solutions)				T_2	(ii) $(01, 41, 13), (10, 33, 12), (32, 21, 02), (11, 24, 20) \bmod (5, 5)$
48	37	111	4	12	1	—	$(00, 01, 12, 15), (01, 03, 08, 10), (0\infty, 07, 15, 21) \bmod (3, 11)$
		[varieties are (x, y) $x = 0, 1, 2$; $y = 0, \dots, 10, \infty$ and $(\infty \infty)$]					$(\infty \infty, 00, 10, 20) \bmod (-, 11); (0\infty, 1\infty, 2\infty, \infty \infty)$
53	27	27	13	13	6	S_1	$[(001, 100, 120, 111, 202, 110, 102, 020, 021, 121, 211, 022, 221)] \bmod (3, 3, 3)$
61	15	42	5	14	4	—	$[(0, 1, 4, 9, 11), (0, 1, 4, 10, 12), (\infty, 0, 1, 2, 7)] \bmod 14$
62	15	35	6	14	5	—	$[(\infty, 00, 10, 11, 12, 14), (\infty, 10, 00, 06, 05, 03), (01, 02, 04, 10, 11, 13), (02, 03, 05, 10, 11, 13), (00, 04, 05, 10, 11, 13)] \bmod (-, 7)$
63	22	77	4	14	2	—	$[(00, 03, 09, 010), (00, 10, 12, 17), (00, 10, 19, 110), (00, 02, 15, 18), (00, 03, 14, 17), (00, 04, 13, 19), (00, 05, 12, 16)] \bmod (-, 11)$
66 ^r	36	84	6	14	2	—	$[(01, 06, 15, 12, 23, 24) \bmod (5, -) + R] \bmod (-, 7)$
		[varieties are (x, y) , $x = 0, \dots, 4$, $y = 0, \dots, 6$, and $(\infty \infty)$]					$[(01, 06, 35, 32, 13, 14) \bmod (5, -) + R] \bmod (-, 7)$
67	43	86	7	14	2	—	$(\infty 0, 01, 06, 15, 12, 23, 24) \bmod (5, 7)$
		[varieties are (x, y) $x = 0, \dots, 4, \infty$, $y = 0, \dots, 6$ and (∞, ∞)]					$(\infty 0, 01, 06, 35, 32, 13, 14) \bmod (5, 7)$
							$(\infty 0, \infty \infty, 00, 10, 20, 30, 40) \bmod (-, 7)$
							$(\infty 0, \infty \infty, 00, 10, 20, 30, 40) \bmod (-, 7)$
							$(\infty 0, \infty 1, \infty 2, \infty 3, \infty 4, \infty 5, \infty 6) \bmod (-, -)$
							$(\infty 0, \infty 1, \infty 2, \infty 3, \infty 4, \infty 5, \infty 6) \bmod (-, -)$
71	11	55	3	15	3	—	$[(0, 1, 3), (0, 1, 5), (0, 2, 7), (0, 1, 8), (0, 3, 5)] \bmod 11$
72	13	39	5	15	5	—	$[(0, 1, 2, 4, 8), (0, 1, 3, 6, 12), (0, 2, 5, 6, 10)] \bmod 13$
73	16	80	3	15	2	—	(i) $(0, 1, 3), (0, 3, 8), (0, 2, 12), (0, 1, 7), (0, 4, 9) \bmod 16$
		(two solutions)				E_2	(ii) $[(10, 11, 12), (10, 12, 15), (00, 07, 10), (01, 06, 10), (02, 05, 10), (03, 04, 10), (01, 07, 10), (02, 06, 10), (03, 05, 10), (00, 10, 14)] \bmod (-, 8)$
74	16	48	5	15	4	—	$[(0, 1, 2, 4, 7), (0, 1, 8, 5, 10), (0, 1, 3, 7, 11)] \bmod 16$
75	16	40	6	15	5	—	$[(0, 1, 3, 5, 9, 12), (0, 1, 2, 3, 6, 12)] \bmod 16$
							PC $(0, 8, 1, 9, 2, 10) \bmod 16$
76	21	35	9	15	6	—	$[(00, 01, 02, 04, 10, 11, 12, 14, 22), (00, 06, 05, 03, 26, 24, 23, 22, 10), (10, 16, 15, 13, 26, 24, 23, 22, 00), (04, 01, 03, 10, 12, 16, 24, 21, 22), (00, 02, 06, 14, 11, 13, 24, 21, 22)] \bmod (-, 7)$
84	61	183	5	15	1	G_2	$[(1, 9, 20, 58, 34), (4, 36, 19, 49, 114), (16, 22, 15, 13, 56)] \bmod 61$

give one replication each. The first expression gives the replication

$$\begin{aligned} &(01, 06, 15, 12, 23, 24), (33, 36, 45, 42, 03, 04) \\ &(11, 16, 25, 22, 33, 34) (41, 46, 05, 02, 12, 14) \\ &(21, 26, 35, 32, 43, 44), (\infty \infty, 00, 10, 20, 30, 40) \end{aligned}$$

and similarly the second expression gives another replication. On cyclic development with respect to the second coordinate the rest of the replications are generated.

Four more solutions, nos. 56, 65, 77 and 79, have been found by trial and error. In the case of no. 77, 26 varieties are represented by (x, y) , $x = 0, \dots, 4$; $y = 0, \dots, 4$ and ∞ . The dicyclic solution in blocks of size 6 is

$$[(\infty, 00, 13, 21, 34, 42), (\infty, 00, 12, 24, 31, 43), (\infty, 00, 10, 20, 30, 40)] \bmod (-, 5)$$

$$[(01, 04, 12, 13, 21, 24), (01, 04, 22, 23, 32, 33)] \bmod (5, 5)$$

For no. 79, the parameters are $v = 31$, $b = 93$, $k = 5$, $r = 15$, $\lambda = 2$. A primitive residue of 31 is 3. Consider the set $(3^0, 3^6, 3^{12}, 3^{18}, 3^{24})$ which is same as $(1, 2, 4, 8, 16)$ with an internal multiplier 2. From this two more sets are generated by successively multiplying by 3, $(3, 6, 12, 24, 17)$ and $(9, 18, 5, 10, 20)$. The cyclic solution (in 3 cycles generating 93 blocks) is

$$[(1, 2, 4, 8, 16), (3, 6, 12, 24, 17), (9, 18, 5, 10, 20)] \bmod 31$$

Similarly the solution for no. 65 with parameters $v = 29$, $b = 58$, $k = 7$, $r = 14$, $\lambda = 3$ is

$$[(1, 7, 16, 20, 23, 24, 25), (3, 21, 19, 2, 11, 14, 17)] \bmod 29$$

For no. 56, represent the varieties by (x, y) , $x = 0, \dots, 5$; $y = 0, \dots, 12$ and ∞ . The solution is

$$\begin{aligned} &[(00, 012, 11, 111, 24, 28), (01, 011, 20, 212, 34, 38)] \bmod (5, 13) \\ &(\infty, 00, 10, 20, 30, 40) \bmod (-, 13) \end{aligned}$$

A simple method of constructing resolvable BIBD for the series $v = s^2$, $b = s(s+1)$, $k = s$, $\lambda = 1$ has been found when $GF(s)$ exists. One need not go through the construction of Theorem 7, although the result of Theorem 7 is interesting from the number-theoretic view point. Let $\alpha_0, \dots, \alpha_{s-1}$ represent the elements of $GF(s)$ and $0, 1, \dots, s-1$, the module $M(s)$ of residue classes mod s and represent the varieties by (x, y) where $x \in M(s)$ and $y \in GF(s)$. Consider the set of pairs (x, y)

$$(0 \alpha_0, 1 \alpha_1, \dots, (s-1) \alpha_{s-1})$$

By cyclic development of y with respect to $GF(s)$ one replication is obtained. The initial set for another replication is obtained by multiplying y in the first set by λ an element of $GF(s)$. Thus.

$$(0 \lambda \alpha_0, 1 \lambda \alpha_1, \dots, (s-1) \lambda \alpha_{s-1})$$

gives on cyclic development of y another replication. Since λ can have s values we obtain s replications. The $(s+1)$ -th replication is obtained by the development with respect to x

$$(0 \alpha_0, 0 \alpha_1, \dots, 0 \alpha_{s-1}) \bmod (s, -).$$

A STUDY OF BIB DESIGNS WITH REPLICATIONS 11 TO 15

For no. 87 represent the varieties by (x, y) , $x = 0, \dots, 6$; $y = 0, \dots, 12$ and ∞ . The solution is

$$(00, 0 \ 12, 14, 18, 51, 511, 26), (33, 39, 42, 4 \ 10, 65, 67, 26) \bmod (7, 13) \\ (00, 10, 20, 30, 40, 50, 60) \bmod 13$$

Note added in proof : I am thankful to Prof. S. S. Shrikhande for pointing out that the solutions for nos. 41, 42, 43, 45, 65, 73, 74, 79 obtained by me also follow from the general series given by Sprott (1954).

REFERENCES

- BOSE, R. C. (1939) : On the construction of balanced incomplete block designs. *Ann. Eugen.*, **9**, 353-399.
- (1942) : The affine analogue of Singer's theorem. *J. Ind. Math. Soc.*, **6**, 1-15.
- BOSE, R. C., CHOWLA, S. and RAO, C. R. (1944) : On the integral order (mod p.) of quadratics x^2+ax+b , with applications to the construction of minimum functions for $GF(p^2)$, and to some number theory results. *Bull. Cal. Math. Ass.*, **36**, 153-174.
- (1945a) : Minimum functions in Galois fields. *Proc. Nat. Acad. Sc.*, **14**, 191.
- (1945b) : On the roots of a well known congruence. *Proc. Nat. Acad. Sc.*, **14**, 193.
- BOSE, R. C. and SHRIKHANDE, S. S. (1960) : On the composition of balanced incomplete block designs. *Can. J. Math.*, **12**, 177-188.
- BRUCK, R. H. and RYSER, H. (1949) : The non-existence of certain finite projective planes. *Can. J. Math.*, **1**, 88-93.
- CARMICHAEL, R. D. (1937) : *Introduction to the Theory of Groups of Finite Order*, Gin and Company.
- CHOWLA, S. and RYSER, H. J. (1950) : Combinatorial problems. *Can. J. Math.*, **2**, 95.
- FISHER, R. A. (1945) : A system of confounding for factors with more than two alternatives, giving completely orthogonal cubes and higher powers. *Ann. Eugen.*, **12**, 283-290.
- FISHER, R. A. and YATES, F. (1953) : *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver and Boyd. Edinburgh.
- HALL, MARSHAL JR. and CONNOR, W. S. (1954) : An embedding theorem for incomplete block designs. *Can. J. Math.*, **6**, 35.
- HALL, MARSHAL JR. (1958) : Some aspects of analysis and probability : *A survey of combinatorial analysis*, 36-106. John Wiley & Sons, New York.
- RAO, C. R. (1944) : Extension of the difference sets of Singer and Bose. *Science and Culture*, **10**, 57.
- (1945) : Finite geometries and certain derived results in theory of numbers. *Proc. Nat. Inst. Sc.* **11**, 136-149.
- (1946) : Difference sets and combinatorial arrangements derivable from finite geometries. *ibid.*, **12**, 123-135.
- SCHÜTZENBERGER, M. P. (1949) : A non-existence theorem for an infinite family of symmetrical block design. *Ann. Eugen.*, **14**, 286-87.
- SHRIKHANDE, S. S. (1950) : The impossibility of certain symmetrical balanced incomplete block designs. *Ann. Math. Stat.*, **21**, 106-11.
- (1960) : Relation between certain incomplete block designs. *Contributions to Probability and Statistics, Essays in Honor of Harold Hotelling*, 388-395. Stanford University Press.
- SPROTT, D. A. (1954) : Note on balanced incomplete block designs. *Can. J. Math.*, **6**, 34-46.
- SKOLEM, TH. (1958) : Some remarks on the triple systems of Steiner. *Math. Scandinavica*, **6**, 273-280.

Paper received : January, 1961.

A TABLE OF MINIMUM FUNCTIONS FOR GENERATING GALOIS FIELDS OF $GF(p^n)$

J. D. ALANEN and D. E. KNUTH

Case Institute of Technology

and

California Institute of Technology

1. MOTIVATION

A polynomial $f(x)$ of degree n irreducible in the field $GF(p)$ where p is a prime number, is called a minimum function, if a root w of the equation $f(x) = 0$, serves as a primitive element of $GF(p^n)$, that is, $w^0 = 1, w, w^2, \dots, w^{p^n-2}$ are the p^n-1 non-zero elements of $GF(p^n)$. It is known that for the $GF(p^n)$, there are exactly $\varphi(p^n-1)/n$ minimum functions, where φ is the Euler function, p a prime, and n an integer.

Minimum functions were used successfully in the past in constructing sets of mutually Orthogonal Latin Squares, Balanced Incomplete Block designs, Confounded and Fractional Factorial designs. Recently, these have found a new application in the construction of error-correcting codes in the theory of information.

While searching for a minimum function of $GF(13^3)$ at the request of Professor C. R. Rao, who needed it for his research work on Designs, we noticed a lack of comprehensive tables of minimum functions in the published literature. The authors are now working on such a set of tables of minimum functions for generating $GF(p^n)$. This will extend considerably the table given in Carmichael (1937).

2. TABLE

Only one minimum function for each one of the cases $p = 11, 13, 17$ and $n = 3, 4, 5$ is given below :

$$GF(11^3) : f(X) = X^3 + X^2 + 3,$$

$$GF(11^4) : f(X) = X^4 + 4X^3 + 2,$$

$$GF(11^5) : f(X) = X^5 + 3X^3 + X^2 + X + 3.$$

$$GF(13^3) : f(X) = X^3 + X^2 + 2,$$

$$GF(13^4) : f(X) = X^4 + X^3 + 3X^2 + 2,$$

$$GF(13^5) : f(X) = X^5 + X^3 + X + 2.$$

$$GF(17^3) : f(X) = X^3 + X + 3,$$

$$GF(17^4) : f(X) = X^4 + 4X^2 + X + 3,$$

$$GF(17^5) : f(X) = X^5 + X^4 + X^3 + 3.$$

REFERENCE

CARMICHAEL, R. D., (1937): *Introduction to the Theory of Groups of Finite Order*, Boston, Ginn and Company.

Paper received : April, 1961.

ANALYSIS OF TWO-WAY DESIGNS

By J. ROY and K. R. SHAH

Indian Statistical Institute

SUMMARY. Methods are given for analysis of general two-way designs with recovery of information from row and column-contrasts, using additivity of plot and treatment effects and the physical fact of randomisation. Column-balanced designs are discussed in particular with a numerical illustration.

1. INTRODUCTION

Designs for two-way elimination of heterogeneity have been considered by various authors: Bose and Kishen (1939), Fisher (1936), Rao (1943, 1946), Yates (1937), and Youden (1937). A general method of analysis under the so-called fixed effects Normal model was given by Shrikhande (1951). The purpose of this paper is to validate Shrikhande's (1951) results and to derive methods of combined estimation after recovery of information from row and column-contrasts using only the assumption of additivity of plot and treatment effects, and the physical fact of randomisation.

To estimate treatment effects, three sets of contrasts namely, row-contrasts, column-contrasts and interaction-contrasts are introduced and their distribution induced by the randomisation considered. These contrasts are all uncorrelated. But since the contrasts in different sets have different variances, it is not possible to combine them effectively unless the ratios of these variances are known. Therefore, best linear unbiased estimates are obtained first from interaction-contrasts only, in Section 3.1. The equations for estimation turn out to be the same as those obtained by Shrikhande under the Normal model.

The equations for combined estimation after recovery of information from row and column-contrasts are given in Section 3.2. Methods for estimating the variance ratios that are required in this problem are given in Sections 4.2 and 4.3. The analysis of variance is shown in Section 4.2. Conditions under which a two-way design compares favourably with the corresponding one-way designs are examined in Section 5 and the relative efficiency-factors worked out.

Though the analysis requires rather heavy computations in the general case, it is shown in Section 6 that if the columns of the design, ignoring rows, form a balanced incomplete block design the analysis is much simpler. The special case where columns are balanced and rows partially balanced is discussed in full and a numerical example is given in Section 8. All derivations are given in Section 7.

2. PRELIMINARIES

2.1. *The additive model.* Suppose there are mn plots or experimental units (eu's) on which a comparative trial involving v treatments is to be carried out. The eu's are arranged in a $m \times n$ two-way classification, so that each eu is determined by a pair of co-ordinates (t, u) $t = 1, 2, \dots, m$; $u = 1, 2, \dots, n$. With the (t, u) -th eu is associated a number x_{tu} to be called the plot effect and we assume that if the k -th treatment is applied on the (t, u) -th eu, the 'yield' would be $x_{tu} + \theta_k$ where the parameter θ_k is to be regarded as the effect of the k -th treatment; $k = 1, 2, \dots, v$. This is the so-called *additive* or *no-interaction* model. The purpose of the experiment is to compare the θ_k 's.

We now define

$$\begin{aligned}\mu &= \Sigma \Sigma x_{tu} / mn, \text{ the general mean,} \\ \sigma_1^2 &= \Sigma \rho_t^2 / (m-1), \text{ the between-row variance,} \quad \dots (2.1) \\ \sigma_2^2 &= \Sigma \gamma_u^2 / (n-1), \text{ the between-column variance,} \\ \sigma_0^2 &= \Sigma \Sigma \eta_{tu}^2 / [(m-1)(n-1)], \text{ the interaction variance,}\end{aligned}$$

where $\rho_t = \frac{1}{n} \Sigma_u x_{tu} - \mu$, $\gamma_u = \frac{1}{m} \Sigma_t x_{tu} - \mu$ and $\eta_{tu} = x_{tu} - \rho_t - \gamma_u - \mu$. We shall write

$$w_i = \sigma_0^2 / \sigma_i^2 \quad i = 1, 2 \quad \dots (2.2)$$

for the ratios of the variances.

2.2. *The design.* The treatments are allocated to the eu's in the following manner. First a two-way design, that is, an arrangement of the v treatments in m rows and n columns is taken. The design is thus completely characterized by the numbers ϵ_{ijk} , $i = 1, 2, \dots, m$; $j = 1, 2, \dots, n$; $k = 1, 2, \dots, v$ where $\epsilon_{ijk} = 1$ if the k -th treatment occurs in the intersection of the i -th row and the j -th column of the design and $\epsilon_{ijk} = 0$, otherwise. The k -th treatment thus occurs in $m_{ki} = \sum_{j=1}^n \epsilon_{ijk}$ positions in the i -th row and in $n_{kj} = \sum_{i=1}^m \epsilon_{ijk}$ positions in the j -th column. We shall restrict ourselves to *equi-replicate* designs, that is to those designs, where each treatment occurs altogether in r positions. Thus $\sum_i m_{ki} = \sum_j n_{kj} = r$ and of course, $\sum_k m_{ki} = n$, $\sum_k n_{kj} = m$. We shall call $\mathbf{M} = ((m_{ki}))$ and $\mathbf{N} = ((n_{kj}))$ the *row incidence-matrix* and the *column incidence-matrix* respectively. The rows and the columns of the design are then allotted to the two ways of classification of the eu's independently and at random.

2.3. *Consequences of randomisation.* Let us denote the yield of the eu corresponding to the i -th row and the j -th column of the design by y_{ij} . The randomisation procedure ensures that

$$E(y_{ij}) = \mu + \sum_{k=1}^v \epsilon_{ijk} \theta_k \quad \dots (2.3)$$

and that

$$\text{cov}(y_{ij}, y_{i'j'}) = \left(\delta_{ii'} - \frac{1}{m} \right) \sigma_1^2 + \left(\delta_{jj'} - \frac{1}{n} \right) \sigma_2^2 + \left(\delta_{ii'} - \frac{1}{m} \right) \left(\delta_{jj'} - \frac{1}{n} \right) \sigma_0^2 \dots \quad (2.4)$$

where $\delta_{ii'}$ is the Kronecker symbol, $\delta_{ii'} = 1(0)$ if $i = i'(i \neq i')$.

2.4. *A linear transformation.* Since the y_{ij} 's are correlated, it is convenient to make a linear transformation and obtain uncorrelated random variables. For this purpose, we use the following definitions. A linear function of the form $l = \sum \sum l_{ij} y_{ij}$ is said to be a *contrast* if $\sum \sum l_{ij} = 0$. A contrast l is said to belong to rows, or simply called a *row-contrast* if $l_{i1} = l_{i2} = \dots = l_{in}$ holds for $i = 1, 2, \dots, m$. Similarly, a contrast l is said to be a *column-contrast* if $l_{1j} = l_{2j} = \dots = l_{mj}$ holds for $j = 1, 2, \dots, n$. A contrast l is said to belong to interaction or simply called an *interaction-contrast* if $\sum_i l_{ij} = 0$ for $j = 1, 2, \dots, n$ and $\sum_j l_{ij} = 0$ for $i = 1, 2, \dots, m$. A contrast l is said to be *normalised* if $\sum \sum l_{ij}^2 = 1$. Two contrasts l and $l' = \sum \sum l'_{ij} y_{ij}$ are said to be *orthogonal* if $\sum \sum l_{ij} l'_{ij} = 0$ holds.

If then we make a linear transformation from y_{ij} 's to (i) $G^* = G/\sqrt{mn}$, where $G = \sum \sum y_{ij}$ is the grand total, (ii) a set of $(m-1)$ mutually orthogonal normalised row-contrasts (iii) a set $(n-1)$ mutually orthogonal normalised column-contrasts, and (iv) a set of $(m-1)(n-1)$ mutually orthogonal normalised interaction-contrasts, it can then be shown as in Section (7.1) that the transformation is orthogonal and that these transformed variables are uncorrelated; the variance of any normalised row-contrast being $n\sigma_1^2$, the variance of any normalised column-contrast being $m\sigma_2^2$ and that of any normalised interaction-contrast being σ_0^2 . Since the expectation of each contrast is a linear function of the θ_k 's, the method of least-squares can be used for purposes of estimation.

2.5. *Notation.* We shall write R_i for the total yield of the i -th row, C_j for that of the j -th column, and T_k for that of the k -th treatment; thus

$$R_i = \sum_{j=1}^n y_{ij}, C_j = \sum_{i=1}^m y_{ij} \text{ and } T_k = \sum_{j=1}^n \sum_{i=1}^m e_{ijk} y_{ij}.$$

We shall use the matrix notations: $\mathbf{R} = (R_1, R_2, \dots, R_m)$ $\mathbf{C} = (C_1, C_2, \dots, C_n)$, $\mathbf{T} = (T_1, T_2, \dots, T_v)$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_v)$. Also, a matrix of the form $p \times q$ with all elements unity will be denoted by $\mathbf{E}_{pq}$.

If $\mathbf{A}$ is a positive semi-definite matrix of form $a \times a$ and rank b , it has b positive latent roots, say α_i , $i = 1, 2, \dots, b$. Let $\boldsymbol{\xi}_i$ of the form $1 \times n$ be a latent vector of $\mathbf{A}$ corresponding to the latent root α_i , $i = 1, 2, \dots, b$ such that $\boldsymbol{\xi}_i \boldsymbol{\xi}_j' = \delta_{ij}$. Then the matrix $\mathbf{A}^* = \sum_{i=1}^b \frac{1}{\alpha_i} \boldsymbol{\xi}_i \boldsymbol{\xi}_i'$ will be called a *pseudo-inverse* of $\mathbf{A}$, in the sense of Rao (1955).

3. ESTIMATION OF TREATMENT EFFECTS

3.1. *Estimation from interaction-contrasts.* Since row-contrasts, column-contrasts and interaction-contrasts have different variances, it is not convenient to use them simultaneously for estimation of treatment effects in an efficient way unless the relative magnitudes of these variances are known. We shall, therefore, consider

first the problem of estimation from interaction-contrasts only. Recovery of information provided by row-contrasts and column-contrasts will be taken up in Section 3.2.

As we have pointed out in Section 2.4, any set of $(m-1) \times (n-1)$ mutually orthogonal normalised interaction-contrasts are mutually uncorrelated and each of them has the same variance σ_0^2 . Also, the expectation of each is a linear function of the θ_k 's. Consequently, the method of least-squares can be used to derive linear unbiased estimators with minimum variance ('best' estimators) of linear functions of treatment effects. As will be shown later in Section 7.2, the method of least-squares gives the equations:

$$\theta K = Q \quad \dots (3.1)$$

where the elements of

$$Q = T - \frac{1}{n} R M' - \frac{1}{m} C N' + \frac{rG}{mn} E_{1v} \quad \dots (3.2)$$

are called the *adjusted yields* of the treatments and

$$K = r I - \frac{1}{n} M M' - \frac{1}{m} N N' + \frac{r^2}{mn} E_{vv} \quad \dots (3.3)$$

will be called the *coefficient-matrix* of the two-way design.

Since $KE_{v1} = 0$, rank $(K) \leq v-1$. A two-way design will be said to be *doubly connected* if its coefficient matrix K is of rank $(v-1)$. In whatever follows, we shall assume that the two-way design is doubly connected.

It is well known from the theory of least-squares (see Rao, 1952) that any linear parametric function of the form $\Theta = \sum_{k=1}^v l_k \theta_k$ with $\sum_{k=1}^v l_k = 0$ admits linear unbiased estimators, and amongst them the one with minimum variance is $T = \sum l_k t_k$ where $t = (t_1, t_2, \dots, t_k)$ is any solution of (3.1). To obtain the variance of T express it in the alternative form $T = \sum_{k=1}^v m_k Q_k$ and then $V(T) = (\sum_{k=1}^v l_k m_k) \sigma_0^2$.

It may be noted that the equations (3.1) are the same as obtained by Shrikhande (1951) from the so-called 'Normal' model. The present approach demonstrates the robustness of this procedure and is aesthetically more satisfying to the authors.

3.2. *Recovery of information from row-contrasts and column-contrasts.* In the previous section, we had simply thrown away the row-contrasts and the column-contrasts. If the ratios $w_i = \sigma_0^2 / \sigma_i^2$, $i = 1, 2$ are known, the method of weighted least-squares can be applied on all the three sets of contrasts simultaneously. If the weight for the normalised interaction-contrasts is taken as unity, the weight for normalised row-contrasts will be w_1/n and that for normalised column-contrasts will be w_2/m . As will be shown in Section 7.2, the method of weighted least-squares now gives the equations :

$$\theta \bar{K} = \bar{Q} \quad \dots (3.4)$$

ANALYSIS OF TWO-WAY DESIGNS

where

$$\bar{Q} = T - \frac{1}{n+\Delta_1} R M' - \frac{1}{m+\Delta_2} C N' + \frac{rG}{mn} \left(1 - \frac{\Delta_1}{n+\Delta_1} - \frac{\Delta_2}{m+\Delta_2} \right) E_{1v} \dots (3.5)$$

$$\bar{K} = rI - \frac{1}{n+\Delta_1} M M' - \frac{1}{m+\Delta_2} N N' + \frac{r^2}{mn} \left(1 - \frac{\Delta_1}{n+\Delta_1} - \frac{\Delta_2}{m+\Delta_2} \right) E_{vv} \dots (3.6)$$

and

$$\Delta_1 = \frac{nw_1}{n-w_1}, \Delta_2 = \frac{mw_2}{m-w_2}. \dots (3.7)$$

Then the best estimator of $\Theta = \sum_{k=1}^v l_k \theta_k$, where $\sum_{k=1}^v l_k = 0$, is given by $\bar{T} = \sum_{k=1}^v l_k \bar{t}_k$ where $\bar{t} = (\bar{t}_1, \bar{t}_2, \dots, \bar{t}_k)$ is any solution of (3.4). The variance of $\bar{T}$ is most easily obtained by writing it in the form $\sum_{k=1}^v \bar{m}_k \bar{Q}_k$ and then $V(\bar{T}) = (\sum_{k=1}^v l_k \bar{m}_k) \sigma_0^2$.

Generally, however, the parameters Δ_1 and Δ_2 would be unknown and estimates D_1 and D_2 for them may have to be substituted. In this case, of course, the 'bestness' of $\bar{T}$ as an estimator of Θ would no longer hold, but conceivably if D_1, D_2 are at all good estimators of Δ_1, Δ_2 ; $\bar{T}$ might yet be better than T . Methods for estimating the Δ 's are given in Section 4.2.

4. ANALYSIS OF VARIANCE

4.1. *Analysis of variance of interaction-contrasts.* To estimate σ_0^2 and to carry out an omnibus test of significance of treatment differences, the analysis of variance is to be done as shown in the following table.

TABLE 4.1. ANALYSIS OF VARIANCE

source	degrees of freedom	sum of squares
(1)	(2)	(3)
rows (unadjusted)	$m-1$	$SS_R^* = \frac{1}{n} \sum_{i=1}^m R_i^2 - \frac{G^2}{mn}$
columns (unadjusted)	$n-1$	$SS_C^* = \frac{1}{m} \sum_{j=1}^n C_j^2 - \frac{G^2}{mn}$
treatments (adjusted for rows and columns)	$v-1$	$SS_{tr} = \sum_{k=1}^v Q_k t_k$
error	$v=(m-1)(n-1)-(v-1)$	$SS_0 = SS_I - SS_{tr}$
interaction	$(m-1)(n-1)$	$SS_I = SS_T - SS_R^* - SS_C^*$
total	$mn-1$	$SS_T = \sum_{i=1}^m \sum_{j=1}^n y_{ij}^2 - \frac{G^2}{mn}$

An unbiased estimator of σ_0^2 is provided by the error mean square $MS_0 = SS_0/v$. To examine the significance of treatment differences, one may use the customary ratio of mean squares MS_{tr}/MS_0 where $MS_{tr} = SS_{tr}/(v-1)$ is the treatment mean square. The sampling distribution of this statistic under randomisation, when the treatment effects are identical, is usually approximated by the Snedecor F -distribution with $(v-1)$ and v degrees of freedom. The accuracy of this approximation is under investigation.

4.2. *Estimation of weights for use in recovery of information.* An estimate of $\Delta_1 = n\sigma_0^2/(n\sigma_1^2 - \sigma_0^2)$ can be obtained as in the case of incomplete block designs by considering the expectation of SS_R , the adjusted row sum of squares. To compute SS_R one has to carry out another analysis of variance ignoring rows. Let $\mathbf{t}_1$ be any solution of the following equations in $\mathbf{0}$

$$\mathbf{0} \mathbf{K}_1 = \mathbf{Q}_1 \quad \dots (4.1)$$

where
$$\mathbf{Q}_1 = T - \frac{1}{m} \mathbf{C} \mathbf{N}' \text{ and } \mathbf{K}_1 = r \mathbf{I} - \frac{1}{m} \mathbf{N} \mathbf{N}'. \quad \dots (4.2)$$

Then
$$SS_R = SS_R^* + SS_{tr} - \mathbf{Q}_1 \mathbf{t}_1'. \quad \dots (4.3)$$

We shall show in Section 7.3 that writing $MS_R = SS_R/(m-1)$ for the adjusted row mean square, the expectation of MS_R is given by

$$E(MS_R) = [1 + (n - a_1)/\Delta_1] \sigma_0^2 \quad \dots (4.4)$$

where
$$a_1 = tr \mathbf{K}_1^* \mathbf{M} \mathbf{M}' / (m-1) \quad \dots (4.5)$$

in which tr denotes the *trace* of a matrix and $\mathbf{K}_1^*$ is a pseudo-inverse of the matrix $\mathbf{K}_1$.

Consequently, we can take

$$D_1 = \frac{(n - a_1)MS_0}{MS_R - MS_0} \quad \dots (4.6)$$

as an estimator of Δ_1 in the sense that the ratio of the expectations of the numerator and the denominator of D_1 is equal to Δ_1 . An estimator D_2 of Δ_2 can be obtained in like manner.

4.3. *A positive-definite estimator of σ_1^2 .* If (4.4) is used for estimation of σ_1^2 , the estimate might on occasions turn out to be negative. We propose here an alternative procedure which is applicable to certain types of designs and has the advantage of providing a positive-definite estimator of σ_1^2 .

As will be shown in Section 7.2, the least-squares equations for estimation of θ_k 's from row-contrasts are

$$\mathbf{0} \left(\mathbf{M} \mathbf{M}' - \frac{r^2}{m} \mathbf{E}_{vv} \right) = \left(\mathbf{R} \mathbf{M}' - \frac{rG}{m} \mathbf{E}_{1v} \right). \quad \dots (4.7)$$

ANALYSIS OF TWO-WAY DESIGNS

If $\mathbf{t}^*$ is any solution of the above equations in $\boldsymbol{\theta}$ and if

$$p = \text{rank} \left(\mathbf{M} \mathbf{M}' - \frac{r^2}{m} \mathbf{E}_{vv} \right) < m-1 \quad \dots \quad (4.8)$$

the residual sum of squares in the analysis of variance of normalised row-contrasts can be used to estimate σ_1^2 . Thus,

$$s_1^2 = \frac{\left(\sum_{i=1}^m R_i^2 - G^2/m \right) - \mathbf{R} \mathbf{M}' \mathbf{t}^*}{n^2(m-1-p)} \quad \dots \quad (4.9)$$

is a positive-definite unbiased estimator of σ_1^2 . The corresponding estimator of Δ_1 is $D_1^* = nMS_0/(ns_1^2 - MS_0)$.

5. EFFICIENCY

It is known [Kempthorne (1956), Roy (1957)] that the average variance of interaction estimators of differences of the type $\theta_k - \theta_{k'}$, is $2\sigma_0^2/h(\mathbf{K})$ where $\mathbf{K}$ is the coefficient-matrix of the design and $h(\mathbf{K})$ denotes the harmonic mean of the positive latent roots of $\mathbf{K}$. If instead of the two-way design, a one-way design using columns as blocks were used, the average variance of intra-block estimates would then be $2 \left[\left(1 - \frac{1}{n} \right) \sigma_0^2 + \sigma_1^2 \right] / h(\mathbf{K}_1)$ where $\mathbf{K}_1 = r\mathbf{I} - \frac{1}{m} \mathbf{N} \mathbf{N}'$. As a measure of the efficiency of the two-way design in comparison with the one-way (column) design, we propose the ratio of the reciprocals of these average variances. This turns out to be

$$E = e\phi \quad \dots \quad (5.1)$$

where $e = h(\mathbf{K})/h(\mathbf{K}_1)$ will be called the *efficiency-factor* of the two-way design relative to the one-way design using columns as blocks and $\phi = 1 + \frac{1}{\Delta_1}$.

It will be shown in Section 7.4 that the relative efficiency-factor $e \leq 1$. Consequently, the two-way design is effective only if $\phi > 1/e$. This parameter ϕ can be estimated by substituting the estimate of Δ_1 as obtained in Section 4.

6. TWO-WAY DESIGNS WITH COLUMN BALANCE

A two-way design will be said to have column balance if each treatment occurs in a column at most once, and any pair of treatments occur together in the same number, say λ , of columns; or in other words if the columns of the design regarded as blocks form a Balanced Incomplete Block Design. A column balanced design is said to be a Youden Square if the row incidence-matrix $\mathbf{M} = \mathbf{E}_{vm}$ and an extended Youden Square if $\mathbf{M} = p\mathbf{E}_{vm}$ where p is a positive integer, $p \geq 2$.

Shrikhande (1951) claims that all known column-balanced designs can be arranged in rows in such a way that (i) a partially balanced association scheme with two associate classes can be imposed on the treatments and (ii) the m_{ki} 's satisfy :

$$\sum_i m_{ki} m_{k'i} = \begin{cases} \mu_0 & \text{if } k = k' \\ \mu_u & \text{if } k \neq k' \end{cases} \text{ are } u\text{-th associates; } u = 1, 2. \quad \dots \quad (6.1)$$

For a definition of a partially balanced association scheme, the reader is referred to Bose and Shimamoto (1952).

If $\mu_1 \neq \mu_2$, the designs satisfying (6.1) are said to belong to the class Y_1 (Shrikhande, 1951).

The analysis of designs belonging to the class Y_1 is particularly easy, being similar to that of partially balanced incomplete block designs with two associate classes. The analysis based on only the interaction-contrasts under the 'Normal' model is given by Shrikhande (1951): here we give the complete analysis including recovery of information from row-contrasts and column-contrasts.

For the parameters of the partially balanced association scheme, we shall use, the standard notations p_{jk}^i , n_i ; $i, j, k = 1, 2$. Let now,

$$\begin{aligned} a &= \frac{\lambda v}{m} - \frac{1}{n} (\mu_0 - \mu_2) \\ b &= \frac{1}{n} (\mu_1 - \mu_2) \quad \dots (6.2) \\ c &= p_{11}^1 - p_{11}^2 \\ d &= n_1 - p_{11}^2 \end{aligned}$$

Then a solution of the equations (3.1) for a design of the class Y_1 turns out to be

$$t_k = [A Q_k + b S_1(Q_k)]/D \quad \dots (6.3)$$

where S_1 denotes summation over first associates and

$$\begin{aligned} A &= a + bc \\ D &= aA - b^2d. \quad \dots (6.4) \end{aligned}$$

Since the variances of estimates of treatment differences are given by

$$V(t_k - t_{k'}) = \begin{cases} [2(A-b)/D] \sigma_0^2, & \text{if } k, k' \text{ are first associates} \\ [2A/D] \sigma_0^2, & \text{otherwise} \end{cases}$$

the relative efficiency factor of this design turns out to be

$$e = \frac{m(v-1)D}{\lambda v[(v-1)A - n_1 b]}. \quad \dots (6.5)$$

For combined estimation, a solution of the equations (3.4) is

$$\hat{t}_k = [\bar{A} \bar{Q}_k + \bar{b} S_1(\bar{Q}_k)]/\bar{D} \quad \dots (6.6)$$

where

$$\begin{aligned} \bar{A} &= \bar{a} + \bar{b}c \\ \bar{D} &= \bar{a}\bar{A} - \bar{b}^2d \quad \dots (6.7) \end{aligned}$$

ANALYSIS OF TWO-WAY DESIGNS

and

$$\bar{a} = \frac{\lambda v + r \Delta_2}{m + \Delta_2} - \frac{1}{n + \Delta_1} (\mu_0 - \mu_2) \quad \dots (6.8)$$

$$\bar{b} = \frac{1}{n + \Delta_1} (\mu_1 - \mu_2).$$

To estimate Δ_1 we need a solution of (4.1) which in this case turns out to be

$$t_{1k} = \frac{m}{\lambda v} Q_{1k}$$

where Q_{1k} is the k -th element of $\mathbf{Q}_1$ defined by (4.2). An estimate of Δ_1 is obtained by putting

$$a_1 = \frac{m\mu_0 - r^2}{\lambda(m-1)} \quad \dots (6.9)$$

in the expression (4.6) for D_1 .

Similarly to estimate Δ_2 we need MS_C the adjusted column mean square given by

$$(n-1) MS_C = SS_C^* + SS_{tr} - \mathbf{Q}_2 \mathbf{t}'_2 \quad (6.10)$$

$\mathbf{t}_2$ being any solution for $\boldsymbol{\theta}$ in

$$\boldsymbol{\theta} \mathbf{K}_2 = \mathbf{Q}_2 \quad \dots (6.11)$$

where $\mathbf{Q}_2 = \mathbf{T} - \frac{1}{n} \mathbf{R} \mathbf{M}'$ and $\mathbf{K}_2 = r \mathbf{I} - \frac{1}{n} \mathbf{M} \mathbf{M}'$ (6.12)

In this case $\mathbf{t}_2 = (t_{21}, \dots, t_{2k})$ is given by

$$\mathbf{t}_{2k} = [A' Q_{2k} + b S[Q_{2k}]] / D' \quad \dots (6.13)$$

where $A' = A + (r - \lambda)/m$ and $D' = [a + (r - \lambda)/m] A' - b^2 d$. An estimate of Δ_2 is given by

$$D_2 = \frac{(m - a_2) MS_0}{MS_C - MS_0} \quad \dots (6.14)$$

where $a_2 = \frac{(r - \lambda)[(v - 1)A' - n_1 b]}{(m - 1)D'}$ (6.15)

The derivation of (6.9) and (6.15) is given in Section 7.5.

7. DERIVATION OF RESULTS

In this section we prove some of the results stated earlier.

7.1. Some lemmas.

Lemma 1: If $l_{ij}^{(\alpha\beta)}$, $i, \alpha = 1, 2, \dots, m$; $j, \beta = 1, 2, \dots, n$ are real numbers chosen to satisfy: (a) $\sum_{ij} l_{ij}^{(\alpha\beta)} l_{ij}^{(\alpha'\beta')} = \delta_{\alpha\alpha'} \delta_{\beta\beta'}$, where δ is the Kronecker symbol,

(b) $l_{ij}^{(mn)} = 1/\sqrt{mn}$, (c) $l_{ij}^{(\alpha m)} = l_{i0}^{(\alpha m)}$, $\alpha = 1, 2, \dots, m-1$ and (d) $l_{ij}^{(m\beta)} = l_{0j}^{(m\beta)}$, $\beta = 1, 2, \dots, n-1$ then we have:

$$\sum_{\alpha=1}^{m-1} l_{ij}^{(\alpha n)} l_{i'j'}^{(\alpha n)} = \frac{1}{n} \left(\delta_{ii'} - \frac{1}{m} \right)$$

$$\sum_{\beta=1}^{n-1} l_{ij}^{(m\beta)} l_{i'j'}^{(m\beta)} = \frac{1}{m} \left(\delta_{jj'} - \frac{1}{n} \right) \quad \dots (7.1)$$

$$\sum_{\alpha=1}^{m-1} \sum_{\beta=1}^{n-1} l_{ij}^{(\alpha\beta)} l_{i'j'}^{(\alpha\beta)} = \left(\delta_{ii'} - \frac{1}{m} \right) \left(\delta_{jj'} - \frac{1}{n} \right)$$

and

$$\sum_{ij} \sum_{i'j'} l_{ij}^{(\alpha\beta)} l_{i'j'}^{(\alpha'\beta')} \delta_{ii'} = \begin{cases} n \delta_{\alpha\alpha'}, & \text{when } \beta = \beta' = n \\ 0 & \text{otherwise} \end{cases}$$

$$\sum_{ij} \sum_{i'j'} l_{ij}^{(\alpha\beta)} l_{i'j'}^{(\alpha'\beta')} \delta_{jj'} = \begin{cases} m \delta_{\beta\beta'}, & \text{when } \alpha = \alpha' = m \\ 0 & \text{otherwise} \end{cases} \quad \dots (7.2)$$

$$\sum_{ij} \sum_{i'j'} l_{ij}^{(\alpha\beta)} l_{i'j'}^{(\alpha'\beta')} \delta_{ii'} \delta_{jj'} = \delta_{\alpha\alpha'} \delta_{\beta\beta'}.$$

These follow easily from the properties of orthogonal matrices.

Lemma 2: Let $Z_{\alpha\beta} = \sum_{ij} l_{ij}^{(\alpha\beta)} y_{ij}$. Then

$$E(Z_{\alpha\beta}) = \begin{cases} \frac{1}{\sqrt{mn}} (mn\mu + r \sum_k \theta_k) & \text{when } \alpha = m, \beta = n \\ \sum_k a_k^{(\alpha\beta)} \theta_k & \text{otherwise} \end{cases} \quad \dots (7.3)$$

where

$$a_k^{(\alpha\beta)} = \sum_{ij} l_{ij}^{(\alpha\beta)} e_{ijk}.$$

Also, these $Z_{\alpha\beta}$'s are mutually uncorrelated and they have variances given by

$$V(z_{\alpha\beta}) = \begin{cases} 0 & \text{if } \alpha = m; \beta = n \\ n\sigma_1^2 & \text{if } \beta = n; \alpha = 1, 2, \dots, m-1 \\ m\sigma_2^2 & \text{if } \alpha = m; \beta = 1, 2, \dots, n-1 \\ \sigma_0^2 & \text{if } \alpha = 1, 2, \dots, m-1; \beta = 1, 2, \dots, n-1. \end{cases} \quad \dots (7.4)$$

These results are obtained by direct computation using the expectations and covariances of y_{ij} 's given by (2.3) and (2.4) and the properties of $l_{ij}^{(\alpha\beta)}$'s given by (7.2).

ANALYSIS OF TWO-WAY DESIGNS

Lemma 3: With $Z_{\alpha\beta}$ and $a_k^{(\alpha\beta)}$'s as defined in Lemma 2, we have

$$Q'_k = \sum_{\alpha=1}^{m-1} Z_{\alpha n} a_k^{(\alpha n)} = \frac{1}{n} \sum_{i=1}^m m_{ki} R_i - \frac{rG}{mn}$$

$$Q''_k = \sum_{\beta=1}^{n-1} Z_{m\beta} a_k^{(m\beta)} = \frac{1}{m} \sum_{j=1}^n n_{kj} C_j - \frac{rG}{mn} \quad \dots \quad (7.5)$$

$$Q_k = \sum_{\alpha=1}^{m-1} \sum_{\beta=1}^{n-1} Z_{\alpha\beta} a_k^{(\alpha\beta)} = T_k - \frac{1}{n} \sum_{i=1}^m m_{ki} R_i - \frac{1}{m} \sum_{j=1}^n n_{kj} C_j + \frac{rG}{mn}$$

$$\gamma'_{kk'} = \sum_{\alpha=1}^{m-1} a_k^{(\alpha n)} a_{k'}^{(\alpha n)} = \frac{1}{n} \sum_{i=1}^m m_{ki} m_{k'i} - \frac{r^2}{mn}$$

$$\gamma''_{kk'} = \sum_{\beta=1}^{n-1} a_k^{(m\beta)} a_{k'}^{(m\beta)} = \frac{1}{m} \sum_{j=1}^n n_{kj} n_{k'j} - \frac{r^2}{mn} \quad \dots \quad (7.6)$$

$$\gamma_{kk'} = \sum_{\alpha=1}^{m-1} \sum_{\beta=1}^{n-1} a_k^{(\alpha\beta)} a_{k'}^{(\alpha\beta)} = r\delta_{kk'} - \frac{1}{n} \sum_{i=1}^m m_{ki} m_{k'i} - \frac{1}{m} \sum_{j=1}^n n_{kj} n_{k'j} + \frac{r^2}{mn}.$$

These results are obtained by direct computation using (7.1) and (7.2).

7.2. *Derivation of least-square equations.* According to the method of weighted least-squares, we have to minimise $\sum \frac{1}{V(Z_{\alpha\beta})} [Z_{\alpha\beta} - E(Z_{\alpha\beta})]^2$, where the summation is over all values of α, β except $\alpha = m, \beta = n$. On multiplication by σ_0^2 , this reduces to minimising

$$L = \sum_{\alpha=1}^{m-1} \sum_{\beta=1}^{n-1} [Z_{\alpha\beta} - E(Z_{\alpha\beta})]^2 + \frac{w_1}{n} \sum_{\alpha=1}^{m-1} [Z_{\alpha n} - E(Z_{\alpha n})]^2 + \frac{w_2}{m} \sum_{\beta=1}^{n-1} [Z_{m\beta} - E(Z_{m\beta})]^2.$$

Equating the partial derivative of L with respect to θ_k ($k = 1, 2, \dots, v$) to zero, we get the equations:

$$\sum_{k'=1}^v \left(\gamma_{kk'} + \frac{w_1}{n} \gamma'_{kk'} + \frac{w_2}{m} \gamma''_{kk'} \right) \theta_{k'} = Q_k + \frac{w_1}{n} Q'_k + \frac{w_2}{m} Q''_k, \quad \dots \quad (7.7)$$

$$k = 1, 2, \dots, v$$

where Q_k and $\gamma_{kk'}$, etc., are given by (7.5) and (7.6). This, in matrix notation is our equation (3.4) for combined estimation. To obtain the equation (3.1) for estimation from interaction-contrasts only, set $w_1 = w_2 = 0$ in (7.7). Similarly if we want estimates from row-contrasts only, the equations would be $\sum_{k'=1}^v \gamma'_{kk'} \theta_{k'} = Q'_k$ which, in matrix notation, is our equation (4.7).

7.3. *Expectation of the adjusted row sum of squares.* Since the adjusted row sum of squares SS_R defined by (4.3) is invariant under the transformation $y'_{ij} = y_{ij} - \sum_{k=1}^v e_{ijk} \theta_k$, its distribution and therefore its expectation does not involve θ_k 's. Consequently in computing $E(SS_R)$ we can ignore the terms involving θ_k 's. Now, since

$$E(SS_R^*) = (m-1)n\sigma_1^2 + \text{terms in } \theta_k\text{'s}$$

and

$$E(SS_{tr}) = (v-1)\sigma_0^2 + \text{terms in } \theta_k\text{'s} \quad \dots (7.8)$$

it follows from (4.3) that all that we need now to compute $E(SS_R)$ is $E(Q_1 t_1')$ where Q_1, t_1 are defined by (4.1) and (4.2). Let K_1^* be a pseudo-inverse of the matrix K_1 defined by (4.2), so that a particular solution of (4.1) is $t_1 = Q_1 K_1^*$. Hence

$$\begin{aligned} E(Q_1 t_1') &= E(Q_1 K_1^* Q_1') = E(\text{tr}[K_1^* Q_1' Q_1]) \\ &= \text{tr}[K_1^* E(Q_1' Q_1)] \\ &= \text{tr} K_1^* D(Q_1) + \text{terms in } \theta_k\text{'s} \quad \dots (7.9) \end{aligned}$$

where $D(Q_1)$ stands for the dispersion matrix of Q_1 . To compute $D(Q_1)$ we express Q_1 in the form $Q_1 = Q + \frac{1}{n} R \left(M' - \frac{r E_{mv}}{m} \right)$. Since the elements of Q are interaction-contrasts and those of $\frac{1}{n} R \left(M' - \frac{r E_{mv}}{m} \right)$ are row-contrasts, these are uncorrelated, and therefore $D(Q_1) = D(Q) + D \left[\frac{1}{n} R \left(M' - \frac{r E_{mv}}{n} \right) \right]$. Since $D(Q) = K \sigma_0^2$ and $D(R) = \left(I - \frac{E}{m} \right) n^2 \sigma_1^2$; we get on simplification

$$D(Q_1) = \left(K_1 + \frac{r^2 E_{vv}}{mn} \right) \sigma_0^2 + MM' \left(\sigma_1^2 - \frac{\sigma_0^2}{n} \right). \quad \dots (7.10)$$

Using (7.8), (7.9) and (7.10), we get finally

$$\begin{aligned} E(SS_R) &= E(SS_R^*) + E(SS_{tr}) - E(Q_1 t_1') \\ &= [n(m-1) - \text{tr} K_1^* MM'] \sigma_1^2 + (\text{tr} K_1^* MM') \sigma_0^2 / n \end{aligned}$$

from which (4.4) follows.

7.4. *Relative efficiency factor.* To prove that the relative efficiency factor e defined in Section 5 cannot exceed unity, we need the following result in matrix theory.

Lemma: If A and B are positive-definite matrices of the same order and $C = A - B$ is positive definite (semi-definite), then $D = B^{-1} - A^{-1}$ is positive definite (semi-definite).

ANALYSIS OF TWO-WAY DESIGNS

Proof: We observe that if $\mathbf{P}$ and $\mathbf{Q}$ are symmetric matrices of the same order and $\mathbf{P}$ is positive definite, a necessary and sufficient condition for $\mathbf{Q}$ to be positive definite (semi-definite) is that the roots of the determinantal equation $|\mathbf{Q} - \lambda \mathbf{P}| = 0$ are all positive (non-negative). Now, the determinantal equation $|\mathbf{C} - \lambda \mathbf{A}| = 0$ is equivalent to the equation $|\mathbf{D} - \lambda \mathbf{B}^{-1}| = 0$. This follows by pre- and post-multiplying the former equation by $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ respectively. Also, $\mathbf{C}$ being positive definite (semi-definite) implies that the root λ are all positive (non-negative) which in turn implies that $\mathbf{D}$ is positive definite (semi-definite).

Next consider a $(v-1) \times v$ matrix $\mathbf{P}$ satisfying $\mathbf{P}\mathbf{P}' = \mathbf{I}$ and $\mathbf{P}'\mathbf{P} = \mathbf{I} - \frac{\mathbf{E}_{vv}}{v}$. Since our design is doubly connected it follows that $\mathbf{A} = \mathbf{P}\mathbf{K}_1\mathbf{P}'$ and $\mathbf{B} = \mathbf{P}\mathbf{K}\mathbf{P}'$ are both positive definite and $\mathbf{A} - \mathbf{B} = \frac{1}{n}\mathbf{P}\mathbf{M}\mathbf{M}'\mathbf{P}'$ is positive definite or semi-definite.

$$\begin{aligned} \text{Now } H(\mathbf{K}_1) - H(\mathbf{K}) &= H(\mathbf{A}) - H(\mathbf{B}) = (v-1) \left(\frac{1}{\text{tr}\mathbf{A}^{-1}} - \frac{1}{\text{tr}\mathbf{B}^{-1}} \right) \\ &= \frac{(v-1)\text{tr}(\mathbf{B}^{-1} - \mathbf{A}^{-1})}{(\text{tr}\mathbf{A}^{-1})(\text{tr}\mathbf{B}^{-1})} \geq 0 \end{aligned}$$

from which it follows that $e \leq 1$.

7.5. Estimation of weights for two-way designs with column-balance.

Since in this case $\mathbf{K}_1 = \frac{\lambda v}{m} \left[\mathbf{I} - \frac{\mathbf{E}}{v} \right]$, we easily get $\mathbf{K}_1^* = \frac{m}{\lambda v} \left[\mathbf{I} - \frac{\mathbf{E}}{v} \right]$. Hence $\mathbf{K}_1^* \mathbf{M}\mathbf{M}' = \frac{m}{\lambda v} \mathbf{M}\mathbf{M}' - \frac{mnr}{\lambda v^2} \mathbf{E}$ from which (6.9) follows very easily using (4.5).

Since $\mathbf{N}\mathbf{N}' = (r-\lambda)\mathbf{I} + \lambda\mathbf{E}$, we have $\text{tr}\mathbf{K}_1^* \mathbf{N}\mathbf{N}' = (r-\lambda) \text{tr}\mathbf{K}_2^*$. Now if $\mathbf{t} = \mathbf{Q}_2\mathbf{Y}$ be any solution in $\boldsymbol{\theta}$ of the equations (6.11), it can be shown that $\text{tr}\mathbf{K}_2^* = \text{tr}(\mathbf{Y}) - \frac{1}{v}(\text{sum of all elements of } \mathbf{Y})$. Since (6.13) provides a solution of the equations (6.11), (6.15) follow immediately.

8. NUMERICAL EXAMPLE

Table 8.1 gives the yields and the lay-out of a design of the Y_1 class with treatments indicated by numbers within brackets.

TABLE 8.1. YIELDS AND THE LAY-OUT AT A DESIGN OF THE Y_1 CLASS (WITH TREATMENTS INDICATED BY NUMBERS WITHIN BRACKETS)

$i \backslash j$	1	2	3	4	5	6	7	8	9	10	R_i
1	140.1 (2)	161.8 (5)	112.2 (1)	153.9 (4)	116.5 (4)	189.2 (6)	160.3 (2)	152.7 (3)	178.0 (3)	134.9 (1)	1499.6
2	102.6 (1)	129.2 (6)	89.5 (3)	97.4 (1)	103.9 (3)	142.5 (5)	138.8 (6)	106.9 (4)	133.3 (5)	87.9 (2)	1132.0
3	155.9 (6)	165.8 (3)	138.3 (6)	141.6 (5)	79.8 (2)	141.6 (4)	161.2 (4)	136.1 (1)	155.8 (2)	107.1 (5)	1383.2
C_j	398.6	456.8	340.0	392.9	300.2	473.3	460.3	395.7	467.1	329.9	4014.8

The parameters of the design are : $m = 3, n = 10, v = 6, r = 5, \lambda = 2,$

$n_1 = 1, n_2 = 4, \mu_0 = 9, \mu_1 = 9, \mu_2 = 8. p_{11}^1 = 0, p_{11}^2 = 0.$

$$a = \frac{\lambda v}{m} - \frac{1}{n}(\mu_0 - \mu_2) = 3.9 \quad b = \frac{1}{n}(\mu_1 - \mu_2) = 0.1$$

$$c = p_{11}^1 - p_{11}^2 = 0 \quad d = n_1 - p_{11}^2 = 1$$

$$A = a + bc = 3.9 \quad D = aA - b^2d = 15.2$$

The computational details of estimation are given in Table 8.2.

TABLE 8.2. COMPUTATIONAL LAY-OUT FOR INTERACTION ESTIMATES AND COMBINED ESTIMATES

treat- ments k	1st asso- ciates of k	T_k	$[R]_k$	$[C]_k$	mnQ_k	$mnDt_k$	mQ_{1k}	nQ_{2k}	$nD't_{2k}$	$\bar{Q}_k$	$\bar{D} \bar{t}_k$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1	3	583.2	6646.4	1857.1	-940.2	-3543.4	-107.5	-814.4	-3965.3	-37.12	-145.29
2	4	623.9	6897.6	1956.1	-1462.8	-5748.9	-84.4	-658.6	-3236.8	-50.04	-202.24
3	1	689.9	6646.4	1959.8	1233.8	4717.8	109.9	252.6	1156.3	39.20	153.82
4	2	680.1	6897.6	2022.4	-439.8	-1861.5	17.9	-96.6	-539.2	-13.45	-58.85
5	6	686.3	6530.0	2120.0	-127.0	-321.7	-61.1	333.0	1730.1	-0.51	3.93
6	5	751.4	6530.0	2129.0	1736.0	6757.7	125.2	984.0	4854.9	61.92	248.63
		4014.8	10G =	3G =	0*	0*	0*	0*	0*	0*	0*
		4014.8*	12044.4*								

$$[R]_k = \sum m_{ki} R_i, [C]_k = \sum n_{kj} C_j;$$

$$mnQ_k = mnT_k - m[R]_k - n[C]_k + rG;$$

$$mnDt_k = A.mnQ_k + bS_1(mnQ_k);$$

$$mQ_{1k} = mT_k - [C]_k;$$

$$nQ_{2k} = nT_k - [R]_k; nD't_{2k} = A'nQ_{2k} + bS_1(nQ_{2k});$$

$$\bar{Q}_k = T_k - \frac{1}{n+D_1}[R]_k - \frac{1}{m+D_2}[C]_k + \frac{r}{mn} \left[1 - \frac{D_1}{n+D_1} - \frac{D_2}{m+D_2} \right]$$

$$\bar{D} \bar{t}_k = \bar{A} \bar{Q}_k + bS_1(\bar{Q}_k).$$

*Denotes check.

If we want interaction estimates only we need proceed only up to column (7) in this table. The analysis of variance table may be prepared at this stage as shown in Table 8.3.

ANALYSIS OF TWO-WAY DESIGNS

TABLE 8.3. ANALYSIS OF VARIANCE

source	degrees of freedom	sum of squares	mean squares	F
rows (unadjusted)	2	$SS_R^* = 7059.34$		
columns (unadjusted)	9	$SS_C^* = 11753.55$		
treatments (adjusted for rows and columns)	5	$SS_{tr} = 2204.15$	$MS_{tr} = 440.83$	3.39
error	13	$SS_0 = 1690.66$	$MS_0 = 130.05$	
interaction	18	$SS_I = 3894.81$		
total	29	$SS_T = 22707.70$		

For recovery of information from row-contrasts and column-contrasts further computations may be made as shown in the columns (8) to (12) of Table 8.2. The constants required are given below.

$$A' = A + (r - \lambda)/m = 4.9, \quad D' = [a + (r - \lambda)/m]A' - b^2d = 24,$$

$$a_1 = \frac{m\mu_0 - r^2}{\lambda(m-1)} = 0.5, \quad a_2 = \frac{(r - \lambda)[(v-1)A' - n_1b]}{(m-1)D'} = 0.3389.$$

We obtain

$$Q_1 t'_1 = 1402.40 \quad Q_2 t'_2 = 4607.75$$

$$SS_R = SS_R^* + SS_{tr} - Q_1 t'_1 = 7861.09 \quad SS_C = SS_C^* + SS_{tr} - Q_2 t'_2 = 9349.95$$

$$D_1 = \frac{(n - a_1)MS_0}{MS_R - MS_0} = 0.3251 \quad D_2 = \frac{(m - a_2)MS_0}{MS_C - MS_0} = 0.3808$$

$$\bar{a} = \frac{\lambda v + rD_2}{m + D_2} - \frac{1}{n + D_1}(\mu_0 - \mu_2) = 4.01579; \quad \bar{b} = \frac{1}{n + D_1}(\mu_1 - \mu_2) = 0.09685$$

$$\bar{A} = \bar{a} + \bar{b}c = 4.01579, \quad \bar{D} = \bar{a}\bar{A} - \bar{b}^2d = 16.11719.$$

The two sets of estimates are given below in Table 8.4.

TABLE 8.4. ESTIMATES OF TREATMENT EFFECTS

k	t_k	$\bar{t}_k$
1	- 7.77	- 9.01
2	-12.61	-12.55
3	10.35	9.54
4	- 4.08	- 3.65
5	- 0.71	0.24
6	14.82	15.43

Estimated variances of the estimates of treatment differences for the two sets are obtained as shown below.

$$\text{Est. } V(t_k - t_{k'}) = \begin{cases} \frac{2(A-b)}{D} \cdot MS_0 = 65.03, & \text{if } k, k' \text{ are first associates} \\ \frac{2A}{D} \cdot MS_0 = 66.74, & \text{otherwise} \end{cases}$$

$$\text{Est. } V(\bar{t}_k - \bar{t}_{k'}) = \begin{cases} \frac{2(\bar{A}-\bar{b})}{\bar{D}} MS_0 = 63.24, & \text{if } (k, k') \text{ are first associates} \\ \frac{2\bar{A}}{\bar{D}} MS_0 = 64.81, & \text{otherwise.} \end{cases}$$

This shows that recovery of information from row-contrasts and column-contrasts has not resulted in appreciable gain in precision.

Next we compare this design with the design formed by taking columns as blocks. The relative efficiency factor e turns out to be 0.97938. Since $\phi = 1 + \frac{1}{\Delta_1} = 4.07598$ we get $E = e \times \phi = 3.99$ as the efficiency of the two-way design relative to the column-design, which shows that the gain in precision is appreciable.

REFERENCES

- BOSE, R. C. and KISHEN, K. (1939): On partially balanced Youden's Squares. *Science and Culture*, **5**, 136-137.
- BOSE, R. C. and SHIMAMOTO, T. (1952): Classification and analysis of partially balanced incomplete block designs with two associate classes. *J. Amer. Stat. Ass.*, **47**, 151-184.
- FISHER, R. A. (1936): *The Design of Experiments*, Oliver and Boyd, Edinburgh.
- KEMPTHORNE, OSCAR (1956): The efficiency factor of an incomplete block design. *Ann. Math. Stat.*, **27**, 846-849.
- RAO, C. R. (1943): Quasi-Latin squares in experimental arrangements. *Current Science*, **7**, 322-323.
- (1946): Confounded factorial designs in quasi-Latin squares. *Sankhyā*, **12**, 295-304.
- (1952): *Advanced Statistical Methods in Biometric Research*, John Wiley and Sons, New York.
- (1955): Analysis of dispersion for multiply classified data with unequal numbers in cells. *Sankhyā*, **15**, 253-280.
- ROY, J. (1958): On the efficiency factor of block designs. *Sankhyā*, **19**, 181-188.
- SHRIKHANDE, S. S. (1951): Designs with two-way elimination heterogeneity. *Ann. Math. Stat.*, **22**, 235-247.
- YATES, F. (1937): The design and analysis of factorial experiments. *Imp. Bur. Soil. Sci. Tech. Comm.*, **35**.
- YODEN, W. J. (1937): Use of incomplete block replications in estimating tobacco-mosaic virus. *Contr. Boyce Thompson Inst.*, **9**, 41-48.

Paper received : November, 1960.

ON SOME PROBLEMS OF FRACTILE GRAPHICAL ANALYSIS

By N. M. MITROFANOVA

Mathematical Institute of the Academy of Sciences, USSR

SUMMARY. Professor Mahalanobis has raised some nonparametric statistical problems in his paper "On fractile graphical analysis" (Mahalanobis, 1959). We consider here some of them under somewhat restrictive conditions on the random variables under consideration.

1. NOTATIONS

Consider any bivariate population of two random variables X and Y with the distribution function $F(x, y)$. Draw two independent samples from this population, each consisting of ng units. Rank these pairs of values of the two variables in the order of ascending values of x .

$$(x_1^*, y_1^*), (x_2^*, y_2^*), \dots, (x_{ng}^*, y_{ng}^*), \quad x_1^* \leq x_2^* \leq \dots \leq x_{ng}^*,$$

$$(x_1'^*, y_1'^*), (x_2'^*, y_2'^*), \dots, (x_{ng}'^*, y_{ng}'^*), \quad x_1'^* \leq x_2'^* \leq \dots \leq x_{ng}'^*$$

and form the functions

$$L(x) = \{L_i(x) = \bar{y}_{i+1}(x-i+1) + \bar{y}_i(i-x), \quad i-1 \leq x \leq i, \quad i = 1, 2, \dots, (g-1) \quad \dots (1)$$

$$L'(x) = \{L_i'(x) = \bar{y}_{i+1}'(x-i+1) + \bar{y}_i'(i-x), \quad i-1 \leq x \leq i, \quad i = 1, 2, \dots, (g-1) \quad \dots (2)$$

where

$$\bar{y}_k = \frac{1}{n} \sum_{j=(k-1)n+1}^{kn} y_j^*, \quad \bar{y}_k' = \frac{1}{n} \sum_{j=(k-1)n+1}^{kn} y_j'^*, \quad k = 1, 2, \dots, g.$$

2. PROBLEM I

Denote

$$a_i = \int_{i-1}^i [\max(L_i, L_i') - \min(L_i, L_i')] dx, \quad i = 1, \dots, (g-1) \quad \dots (3)$$

$$a = \int_0^{g-1} [\max(L, L') - \min(L, L')] dx = \sum_{i=1}^{g-1} a_i. \quad \dots (4)$$

In the terms of the paper mentioned above this is an error area $a(1, 2)$ to be associated with the combined fractile graph $G(1, 2)$.

First of all we shall be interested in Ea and Da ,

3. CASE (i) : X AND Y ARE INDEPENDENT

Let the random variables X and Y be independent, so that $F(x, y) = \Phi(x) \cdot F(y)$, $\Phi(x)$ and $F(y)$ being the distribution functions of X and Y respectively. The relation

$$\int_{x_1^* \leq x_2^* \leq \dots \leq x_{ng}^*} (ng)! d\Phi(x_1^*) dF(y_1^*) \dots d\Phi(x_{ng}^*) dF(y_{ng}^*) = dF(y_1^*) \dots dF(y_n^*),$$

shows that the variables $Y_1^*, Y_2^*, \dots, Y_{ng}^*$ are independent, and $P(Y_i^* < y) = F(y)$, $i = 1, \dots, ng$. The random variables $Y_1'', Y_2'', \dots, Y_{ng}''$ have the same distribution functions and are also independent.

Let $EY = m$ and $DY = \sigma^2$. Since $\max(L_i, L'_i) - \min(L_i, L'_i) = \max(L_i - m, L'_i - m) - \min(L_i - m, L'_i - m)$ we have the right to suppose further $m = 0$.

Evaluation of Ea .

$$Ea = \sum_{i=1}^{g-1} Ea_i = \sum_{i=1}^{g-1} \int_{i-1}^i [E \max(L_i, L'_i) - E \min(L_i, L'_i)] dx. \quad \dots (5)$$

$$\text{Consider} \quad E \max(L_i, L'_i) = T_i. \quad \dots (6)$$

The random variables L_i and L'_i are identically distributed, therefore,

$$T_i = \int z d[P(L_i < z)]^2.$$

$$\text{Making the substitution} \quad z = \frac{\sigma u}{\sqrt{n}} \sqrt{(x-i+1)^2 + (i-x)^2}$$

$$\text{we get} \quad T_i = \frac{\sigma}{\sqrt{n}} \sqrt{(x-i+1)^2 + (i-x)^2} \int u d[G_n(u)]^2 \quad \dots (7)$$

$$\text{where} \quad G_n(u) = P \left(\frac{\sqrt{n}}{\sigma \sqrt{(x-i+1)^2 + (i-x)^2}} L_i < u \right).$$

According to Liapunov's theorem

$$G_n(u) \rightarrow G(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{t^2}{2}} dt \text{ as } n \rightarrow \infty.$$

$$\text{Hence} \quad \frac{\sqrt{n} T_i}{\sigma} \rightarrow \sqrt{(x-i+1)^2 + (i-x)^2} \int u d[G(u)]^2. \quad \dots (8)$$

On the other hand, obviously

$$\min(\bar{y}_i, \bar{y}'_i, \bar{y}_{i+1}, \bar{y}'_{i+1}) \leq \max(L_i, L'_i) \leq \max(\bar{y}_i, \bar{y}'_i, \bar{y}_{i+1}, \bar{y}'_{i+1}).$$

Hence
$$|T_i| \leq \int |z| d[P(\bar{y}_i < z)]^4 \leq 4 \int |z| d[P(\bar{y}_i < z)]$$

$$\leq 4 \sqrt{\int z^2 d[P(\bar{y}_i < z)]} = 4 \sqrt{D \bar{y}_i} = 4 \frac{\sigma}{\sqrt{n}}.$$

or
$$\left| \frac{\sqrt{n} T_i}{\sigma} \right| \leq 4. \quad \dots (9)$$

(8) and (9) enables us to assert that

$$\int_{i-1}^i \frac{\sqrt{n} T_i}{\sigma} dx \rightarrow \int_0^1 \sqrt{x^2 + (1-x)^2} dx. \int ud[G(u)]^2 \text{ as } n \rightarrow \infty.$$

If
$$I_i = E \min(L_i, L'_i)$$

then we would find

$$- \int_{i-1}^i \frac{\sqrt{n} I_i}{\sigma} dx \rightarrow \int_0^1 \sqrt{x^2 + (1-x)^2} dx. \int ud[1-G(u)]^2 \text{ as } n \rightarrow \infty.$$

After some obvious simplifications finally we obtain

$$Ea = (g-1) \frac{\sigma}{\sqrt{n}} \frac{2}{\sqrt{\pi}} \int_0^1 \sqrt{x^2 + (1-x)^2} dx. [1+o(1)]. \quad \dots (10)$$

Evaluation of Da. It is easy to verify that

$$|a_i| \leq \max_k (|\bar{y}_k - \bar{y}'_k|)$$

and thus
$$|a| \leq (g-1) \max (|\bar{y}_k - \bar{y}'_k|).$$

Denote
$$Z_k = |\bar{y}_k - \bar{y}'_k|.$$

Since all the Z_k are independent and have the same distribution functions $H(z)$

$$\begin{aligned} Ea^2 &\leq (g-1)^2 \int z^2 d[H(z)]^g \leq g(g-1)^2 \int z^2 dH(z) \\ &= g(g-1)^2 E Z_k^2 = g(g-1)^2 E(\bar{y}_k - \bar{y}'_k)^2 = g(g-1)^2 \frac{2\sigma^2}{n}. \quad \dots (11) \end{aligned}$$

Hence
$$D(a) = O\left(\frac{\sigma^2}{n}\right). \quad \dots (12)$$

4. CASE (ii) : X AND Y ARE BIVARIATE NORMAL

Let a bivariate random variable have the density function

$$p(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2} \frac{1}{1-\rho^2} \left[\frac{(x-m_1)^2}{\sigma_1^2} - \frac{2\rho(x-m_1)(y-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2} \right] \right\}.$$

Denote
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp \left\{ -\frac{1}{2} \frac{(x-m_1)^2}{\sigma_1^2} \right\}$$

$$p(y | x) = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2} \frac{1}{\sigma_2^2(1-\rho^2)} \left[y-m_2 - \frac{\sigma_2\rho}{\sigma_1}(x-m_1) \right]^2 \right\}.$$

Further, we shall keep $x_1^*, \dots, x_{ng}^*$ and $x_1'^*, \dots, x_{ng}'^*$ fixed with

$$\left. \begin{array}{l} x_1^* \leq \dots \leq x_n^* \\ x_1'^* \leq \dots \leq x_{ng}'^* \end{array} \right\} = x^*.$$

Then the conditional density functions for $Y_1^*, Y_2^*, \dots, Y_{ng}^*$ and $Y_1'^*, Y_2'^*, \dots, Y_{ng}'^*$

will be of the form $\frac{(ng)! p(x_1^*, y_1^*) p(x_2^*, y_2^*) \dots p(x_{ng}^*, y_{ng}^*)}{(ng)! p(x_1^*) p(x_2^*) \dots p(x_{ng}^*)} = p(y_1^* | x_1^*) \dots p(y_{ng}^* | x_{ng}^*)$ and

$p(y_1'^* | x_1'^*) \dots p(y_{ng}'^* | x_{ng}'^*)$ respectively.

Thus we have two series of independent variables. For the sake of shortness we shall denote them by the former letters Y_i^* and $Y_i'^*$, $i = 1, \dots, ng$ with their new sense in mind.

Evaluation of Ea . As before

$$\begin{aligned} |a_i| &\leq \max_k |\bar{y}_k - \bar{y}_k'| = \max_k Z_k, \\ |Ea| &\leq (g-1)E \max_k Z_k. \end{aligned} \quad \dots (13)$$

All the Z_k are independent; let $H_k(z)$ be the distribution function of Z_k . Then

$$\begin{aligned} E \max_k Z_k &= \sum_{i=1}^g \int z \prod_{k \neq i} H_k(z) dH_i(z) \leq \sum_{i=1}^g \int z dH_i(z) \\ &\leq \sum_{i=1}^g \sqrt{\int z^2 dH_i(z)} \leq \sum_{i=1}^g \sqrt{E(\bar{y}_i - \bar{y}_i')^2}. \end{aligned} \quad \dots (14)$$

Let us evaluate the last expected value.

$$E \cdot (\bar{y}_i - \bar{y}_i')^2 = D_{x^*} (\bar{y}_i - \bar{y}_i') + [E_{x^*} (\bar{y}_i - \bar{y}_i')]^2.$$

It is easy to see that

$$\bar{y}_i \in N \left(m_2 + \frac{\sigma_2\rho}{\sigma_1} (\bar{x}_i - m_1), \frac{\sigma_2\sqrt{1-\rho^2}}{\sqrt{n}} \right)$$

and

$$\bar{y}_i' \in N \left(m_2 + \frac{\sigma_2\rho}{\sigma_1} (\bar{x}_i' - m_1), \frac{\sigma_2\sqrt{1-\rho^2}}{\sqrt{n}} \right)$$

where

$$\bar{x}_i = \frac{1}{n} \sum_{j=(i-1)n+1}^{in} x_j^*; \quad \bar{x}_i' = \frac{1}{n} \sum_{j=(i-1)n+1}^{in} x_j'^*.$$

ON SOME PROBLEMS OF FRACTILE GRAPHICAL ANALYSIS

Therefore

$$E_{x^*}(\bar{y}_i - \bar{y}'_i)^2 = \frac{2\sigma_2^2(1-\rho^2)}{n} + \frac{\sigma_2^2\rho^2}{\sigma_1^2} [(\bar{x}_i - m_1) - (\bar{x}'_i - m_1)]^2$$

and the unconditional expected value is

$$E(\bar{y}_i - \bar{y}'_i)^2 = \frac{2\sigma_2^2(1-\rho^2)}{n} + \frac{2\sigma_2^2\rho^2}{\sigma_1^2} D(\bar{x}_i - m_1). \quad \dots (15)$$

Let E' denote the conditional expected value when

$$x_1^* \leq \dots \leq x_{(i-1)n}^* \leq x_{in+1}^* \leq \dots \leq x_{ng}^*$$

are fixed. The conditional density will be

$$\frac{n! p(x_{(i-1)n+1}^*) \dots p(x_{in}^*)}{\left(\int_{x_{(i-1)n}^*}^{x_{in+1}^*} p(x) dx \right)^n}.$$

Then

$$\begin{aligned} E' \left(\sum_{(i-1)n+1}^{in} (x_j^* - m_1) \right)^2 &= \frac{n!}{\left(\int_{x_{(i-1)n}^*}^{x_{in+1}^*} p(x) dx \right)^n} \int_{x_{(i-1)n}^*}^{x_{in+1}^*} \dots \int_{x_{(i-1)n}^*}^{x_{in+1}^*} \left(\sum_{(i-1)n+1}^{in} (x_j^* - m_1) \right)^2 \prod_{(i-1)n+1}^{in} p(x_j^*) dx_j^* \\ &= \frac{1}{\left(\int_{x_{(i-1)n}^*}^{x_{in+1}^*} p(x) dx \right)^n} \int_{x_{(i-1)n}^*}^{x_{in+1}^*} \dots \int_{x_{(i-1)n}^*}^{x_{in+1}^*} \left(\sum_{(i-1)n+1}^{in} (x_j - m_1) \right)^2 \prod_{(i-1)n+1}^{in} p(x_j) dx_j \\ &= \frac{n}{\int_{x_{(i-1)n}^*}^{x_{in+1}^*} p(x) dx} \int_{x_{(i-1)n}^*}^{x_{in+1}^*} (x - m_1)^2 p(x) dx + (n^2 - n) \left[\frac{\int_{x_{(i-1)n}^*}^{x_{in+1}^*} (x - m_1) p(x) dx}{\int_{x_{(i-1)n}^*}^{x_{in+1}^*} p(x) dx} \right]^2. \end{aligned}$$

And similarly

$$E' \left(\sum_{(i-1)n+1}^{in} (x_j^* - m_1) \right) = \frac{n}{\int_{x_{(i-1)n}^*}^{x_{in+1}^*} p(x) dx} \int_{x_{(i-1)n}^*}^{x_{in+1}^*} (x - m_1) p(x) dx.$$

The conditional variance will be equal to

$$D' \left(\sum_{(i-1)n+1}^{in} (x_j^* - m_1) \right) = n \frac{\int_{x_{(i-1)n}^*}^{x_{in+1}^*} (x - m_1)^2 p(x) dx}{\int_{x_{(i-1)n}^*}^{x_{in+1}^*} p(x) dx} - n \left[\frac{\int_{x_{(i-1)n}^*}^{x_{in+1}^*} (x - m_1) p(x) dx}{\int_{x_{(i-1)n}^*}^{x_{in+1}^*} p(x) dx} \right]^2$$

or

$$D' \left(\sum_{(i-1)n+1}^{in} (x_j^* - m_1) \right) = \sigma_1^2 n \left\{ \frac{\int_{x_{(i-1)n}^*}^{x_{in+1}^*} \left(\frac{x - m_1}{\sigma_1} \right)^2 p(x) dx}{\int_{x_{(i-1)n}^*}^{x_{in+1}^*} p(x) dx} - \left[\frac{\int_{x_{(i-1)n}^*}^{x_{in+1}^*} \frac{x - m_1}{\sigma_1} p(x) dx}{\int_{x_{(i-1)n}^*}^{x_{in+1}^*} p(x) dx} \right]^2 \right\} \quad \dots (16)$$

where the expression in braces does not exceed

$$\frac{2}{\left(\int_{x_{(i-1)n}^*}^{x_{in+1}^*} p(x) dx \right)^2} \quad \dots (17)$$

(16) and (17) imply that

$$D \left(\sum_{(i-1)n+1}^{in} (x_j^* - m_1) \right) \leq \sigma_1^2 n \cdot 2 \frac{(ng)!}{n!} \int_{x_1^* \leq \dots \leq x_{(i-1)n}^* \leq x_{in+1}^* \leq \dots \leq x_{ng}^*} \prod_1^{(i-1)n} p(x_j^*) dx_j^* \\ \left(\int_{x_{(i-1)n}^*}^{x_{in+1}^*} p(x) dx \right)^{n-2} \prod_{in+1}^{ng} p(x_j^*) dx_j^* \\ \leq \sigma_1^2 n \cdot 2 \frac{(ng)!}{n!} \frac{(n-2)!}{(ng-2)!} \leq \text{const } g^2 \sigma_1^2 n. \quad \dots (18)$$

$$\text{From (13), (14), (15) and (18) we have } Ea = O \left(\frac{\sigma_2}{\sqrt{n}} \right). \quad \dots (19)$$

Evaluation of D(a). We have

$$a^2 \leq (g-1) \sum_{i=1}^{g-1} a_i^2 \leq (g-1)^2 (\max_k Z_k)^2 \\ E (\max_k Z_k)^2 = \sum_{i=1}^g \int z^2 \prod_{k \neq i} H_k(z) dH_i(z) \leq \sum_{i=1}^g \int z^2 dH_i(z) = \sum_{i=1}^g E(\bar{y}_i - \bar{y}'_i)^2.$$

ON SOME PROBLEMS OF FRACTILE GRAPHICAL ANALYSIS

Thus

$$Da = O\left(\frac{\sigma_z^2}{n}\right) \quad \dots (20)$$

and the first hypothesis of Mahalanobis (1959) is true.

5. PROBLEM II

Let us mix together the two samples we spoke of at the beginning of this note to form a combined sample and again rank the sample-units in the order of ascending values of x . Divide the new series of y 's into g groups of equal number $2n$ as we did before and construct the function

$$M = \{M_i = \bar{Z}_{i+1}(x-i+1) + \bar{Z}_i(i-x), \quad i-1 \leq x \leq i, \quad i = 1, 2, \dots, (g-1),$$

$\bar{Z}_i$ ($i = 1, 2, \dots, g$) being the mean value of y 's in the i -th group.

The second hypothesis of P. C. Mahalanobis says that the graph M "would tend statistically to lie more and more within the area bounded by the two graphs" L and L' . To confirm this hypothesis to be true in the case of normal correlation between X and Y we shall show that the expected length of the projection of the part of this graph which lies outside the area bounded by the graphs L and L' tends to zero as $n \rightarrow \infty$ for constant g . For simplicity we suppose $g = 2$ and first consider the case of independent X and Y .

$$\text{Let } X \in N(0, 1), Y \in N(0, 1) \text{ and } \bar{Z}_1 = \frac{1}{2n} \left(\sum_1^k y_i^* + \sum_1^{2n-k} y_i'^* \right), \bar{Z}_2 = \frac{1}{2n} \left(\sum_{k+1}^{2n} y_i^* + \sum_{2n-k+1}^{2n} y_i'^* \right).$$

(The sense is clear). Such a method of constructing $\bar{Z}_1$ and $\bar{Z}_2$ is equivalent to the event

$$A_k = \left\{ \begin{array}{c} x_k^* < x_{2n-k}^* < x_{k+1}^* \\ \text{or} \\ x_{2n-k}^* < x_k^* < x_{2n-k+1}^* \end{array} \right\}.$$

It is easy to verify, that $P(A_k) = \frac{[C_{2n}^k]^2}{C_{4n}^{2n}}.$

Let
$$\left. \begin{array}{l} x_1^* \leq \dots \leq x_{2n}^* \\ x_1'^* \leq \dots \leq x_{2n}'^* \end{array} \right\} = x^* \quad \text{be fixed.}$$

We are interested in the mean length of the projection (inside the interval $[0, 1]$) of the part of the line $y = (1-x)\bar{Z}_1 + x\bar{Z}_2$ which lies outside the area a .

$$El = E E(l|x^*) \leq 2 \sum_{k=0}^n \int_{A_k} E(l|x^*) dP(x^*)$$

as in any case

$$0 \leq l \leq 1$$

$$\int_{A_k} E(l|x^*) dP(x^*) \leq P(A_k)$$

and we get
$$El \leq 2 \sum_{k=0}^{n-\sqrt{n \ln n}} P(A_k) + 2 \sum_{k=n-\sqrt{n \ln n}+1}^n \int_{A_k} E(l|x^*) dP(x^*). \quad \dots (21)$$

Further, we have
$$\sum_{k=0}^{n-\sqrt{n \ln n}} P(A_k) \leq \frac{2^{4n}}{C_{2n}^{2n}} \left(\sum_0^{n-\sqrt{n \ln n}} C_{2n}^k \frac{1}{2^{2n}} \right)^2;$$

$$\begin{aligned} \sum_0^{n-\sqrt{n \ln n}} C_{2n}^k \frac{1}{2^{2n}} &\rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{2n}}^{-\sqrt{2 \ln n}} e^{-t^2/2} dt \\ &\leq \frac{1}{\sqrt{2\pi}} \int_{\sqrt{2 \ln n}}^{\infty} e^{-t^2/2} dt = O(\exp -\frac{1}{2}(\sqrt{2 \ln n})^2) = O\left(\frac{1}{n}\right). \end{aligned}$$

Thus
$$\sum_{k=0}^{n-\sqrt{n \ln n}} P(A_k) = O\left(\frac{1}{n^{3/2}}\right) \rightarrow 0. \quad \dots (22)$$

Let
$$n - \sqrt{n \ln n} \leq k \leq n$$

and
$$\bar{Z}_1 = \frac{\bar{y}_1 + \bar{y}'_1}{2} - \frac{\sum_{k+1}^n y_i^*}{2n} + \frac{\sum_{n+1}^{2n-k} y_i^*}{2n} = \frac{\bar{y}_1 + \bar{y}'_1}{2} + \delta$$

$$\bar{Z}_2 = \frac{\bar{y}_2 + \bar{y}'_2}{2} + \frac{\sum_{k+1}^n y_i^*}{2n} - \frac{\sum_{n+1}^{2n-k} y_i^*}{2n} = \frac{\bar{y}_2 + \bar{y}'_2}{2} - \delta.$$

Note that $\delta \in N\left(0, C \frac{(\ln n)^{1/4}}{n^{3/4}}\right)$. (The letter C , the same for all cases, will denote various constants inessential for us).

If $X \in N\left(0, C \frac{1}{n^\frac{1}{2}}\right)$ and $\epsilon \geq 0$ is a small number then

$$\begin{aligned} &P\left(|\delta| > \frac{1}{n^{1/8}} |X|^{1+\epsilon}\right) \\ &= P\left(C \frac{(\ln n)^{1/4}}{n^{3/4}} |\xi| > C \frac{1}{n^{1/8}} \frac{1}{n^{(1+\epsilon)/2}} |\eta|^{1+\epsilon}\right) = P\left(C \frac{(\ln n)^{1/4}}{n^{1/8-\epsilon/2}} |\xi| > |\eta|^{1+\epsilon}\right) \\ &= P\left(C \frac{(\ln n)^{1/4}}{n^{1/8-\epsilon/2}} |\xi| > |\eta|^{1+\epsilon}, |\eta| < C \frac{(\ln n)^{\frac{1}{8(1+\epsilon)}}}{\frac{1-4\epsilon}{16(1+\epsilon)}}\right) + \\ &\quad + P\left(C \frac{(\ln n)^{1/4}}{n^{1/8-\epsilon/2}} |\xi| > |\eta|^{1+\epsilon}, |\eta|^{1+\epsilon} > C \frac{(\ln n)^{1/8}}{\frac{1-4\epsilon}{16}}\right) \\ &\leq P\left(|\eta| < C \frac{(\ln n)^{1/8(1+\epsilon)}}{\frac{1-4\epsilon}{16(1+\epsilon)}}\right) + P\left(C \frac{(\ln n)^{1/8}}{\frac{1-4\epsilon}{16}} |\xi| > 1\right) \\ &\leq C \frac{(\ln n)^{\frac{1}{8(1+\epsilon)}}}{n^{(1-4\epsilon)/16(1+\epsilon)}} + C \frac{(\ln n)^{1/8}}{n^{(1-4\epsilon)/16}} E|\xi|. \end{aligned}$$

Here $\xi, \eta \in N(0, 1)$. Thus we have the following lemma.

ON SOME PROBLEMS OF FRACTILE GRAPHICAL ANALYSIS

Lemma: Under the conditions mentioned above we have

$$P\left(|\delta| > \frac{1}{n^{1/8}} |X|^{1+\varepsilon}\right) = O\left(\frac{(\ln n)^a}{n^b}\right) \quad \dots (23)$$

for the proper positive constants a and b .

Now let us investigate $E(l|x^*)$ for $n - \sqrt{n \ln n} \leq k \leq n$.

Let

$$W_1 = \frac{\bar{y}_1 - \bar{y}'_1}{2} \text{ and } W_2 = \frac{\bar{y}_2 - \bar{y}'_2}{2}.$$

According to our lemma

$$P\left(|\delta| > |W_1|, \text{ or } |\delta| > |W_2|, \text{ or } |\delta| > \frac{1}{n^{1/8}} |W_1 - W_2|^{1+\varepsilon} \text{ or } |\delta| > \frac{1}{4} |W_1 - W_2|\right) = O\left(\frac{(\ln n)^a}{n^b}\right), \quad \dots (24)$$

$$\begin{aligned} E(l|x^*) &= \int_{\substack{|\delta| > |W_1| \\ \text{or } |\delta| > |W_2| \\ \text{or } |\delta| > \frac{1}{n^{1/8}} |W_1 - W_2|^{1+\varepsilon} \\ \text{or } |\delta| > \frac{1}{4} |W_1 - W_2|}} l dF(l) + \int_{\substack{|\delta| < |W_1| \\ |\delta| < |W_2| \\ |\delta| < \frac{1}{n^{1/8}} |W_1 - W_2|^{1+\varepsilon} \\ |\delta| < \frac{1}{4} |W_1 - W_2|}} l dF(l) \\ &= O\left(\frac{(\ln n)^a}{n^b}\right) + \int_{\substack{|\delta| < |W_1| \\ |\delta| < |W_2| \\ |\delta| < \frac{1}{n^{1/8}} |W_1 - W_2|^{1+\varepsilon} \\ |\delta| < \frac{1}{4} |W_1 - W_2|}} l dF(l). \quad \dots (25) \end{aligned}$$

When $|\delta| < |W_1|$, $|\delta| < |W_2|$ we have two points of intersection whose abscissae $\bar{x}_1$

$$\text{and } \bar{x}_2 \text{ are given by } \bar{x}_1 = \frac{W_1 - \delta}{W_1 - W_2 - 2\delta} \text{ and } \bar{x}_2 = \frac{W_1 + \delta}{W_1 - W_2 + 2\delta}$$

so that

$$\bar{x}_1 - \bar{x}_2 = \frac{2\delta(W_1 + W_2)}{(W_1 - W_2)^2 - 4\delta^2}$$

$$\text{Therefore we have } l = |\bar{x}_1 - \bar{x}_2| \leq C \frac{|\delta| |W_1 + W_2|}{|(W_1 - W_2)^2 - 4\delta^2|}.$$

$$\leq C \frac{1}{n^{1/8}} \frac{|W_1 + W_2|}{|W_1 - W_2|^{1-\varepsilon}}.$$

It is not difficult to verify that $W_1 + W_2$ and $W_1 - W_2$ are independent. Consider

$u = \sqrt{n}(W_1 + W_2)$ and $v = \sqrt{n}(W_1 - W_2)$ then $u, v \in N(0, 1)$ and $l \leq C \frac{1}{n^{1/8}} |u|^\varepsilon$

$\left(\frac{|u|}{|v|}\right)^{1-\varepsilon}$. The random variable u/v has Cauchy distribution with the density

$\frac{1}{\pi(1+x^2)}$ and hence $E\left|\frac{u}{v}\right|^{1-\varepsilon}$ exists for any ε , $0 < \varepsilon < 1$.

Now let us choose $\Delta > 1$ in such a way that $(1-\epsilon)\Delta = 1-\epsilon_1, 0 < \epsilon_1 < 1$.

According to Holder's inequality for $\frac{1}{\Delta} + \frac{1}{\frac{\Delta}{\Delta-1}} = 1$

$$E \left(\left| u \right|^\epsilon \left| \frac{u}{v} \right|^{1-\epsilon} \right) \leq \left[E \left| \frac{u}{v} \right|^{1-\epsilon_1} \right]^{\frac{1}{\Delta}} \left[E \left| u \right|^{\frac{\epsilon \Delta}{\Delta-1}} \right]^{\frac{\Delta-1}{\Delta}}.$$

The existence of $E \left| u \right|^{\frac{\epsilon \Delta}{\Delta-1}}$ is trivial. Thus

$$\int l dF(l) \leq C \frac{1}{n^{1/8}} \text{ for } n \geq N. \quad \dots (26)$$

$$\begin{aligned} |\delta| &< |W_1| \\ |\delta| &< |W_2| \\ |\delta| &< \frac{1}{n^{1/8}} |W_1 - W_2|^{1+\epsilon} \\ |\delta| &< \frac{1}{4} |W_1 - W_2| \end{aligned}$$

Thus from (25) and (26) we have $E(l|x^*) = O \left(\frac{(\ln n)^a}{n^b} \right)$ for all x^* (27)

$$\max_k P(A_k) = \frac{[C_{2n}^n]^2}{C_{4n}^{2n}} = O \left(\frac{1}{\sqrt{n}} \right) \quad \dots (28)$$

Let
$$\sum_{k=n-\sqrt{n \ln n}+1}^n \int_{A_k} E(l|x^*) dP(x^*) = \Sigma. \quad \dots (29)$$

then
$$\Sigma \leq O \left(\frac{(\ln n)^a}{n^b} \right) \left\{ \max_k P_k \right\} \sqrt{n \ln n} \quad \dots (30)$$

Thus (27), (28) and (30) show $\Sigma = O \left(\frac{(\ln n)^{a+\frac{1}{2}}}{n^b} \right) \rightarrow 0. \quad \dots (31)$

(21), (22) and (31) prove our assertion.

Now it is clear how to paraphrase the above proof for the normal correlation between X and Y . In this case we would have

$$\delta \epsilon N \left(\frac{\rho}{2n} \left(\sum_{i=1}^{2n-k} x_i'^* - \sum_{i=1}^n x_i^* \right), C \frac{(\ln n)^{1/4}}{n^{3/4}} \right)$$

and the Lemma is true for $X = W_1 - W_2$. Then though the density function for $\frac{u}{v}$ will not be of Cauchy type, it is not difficult to prove that $E \left| \frac{u}{v} \right|^{1-\epsilon}$ still exists.

REFERENCE

MAHALANOBIS, P. C. (1959) : On fractile graphical analysis. Indian Statistical Institute : Draft proof, April 1959.

Paper received : November, 1960.

SOME REMARKS CONCERNING THE EXPECTATION OF THE ERROR AREA IN FRACTILE ANALYSIS

By YUKIYOSI KAWADA

University of Tokyo

SUMMARY. Expressions for the expectation of the error area in Fractile Graphical Analysis with samples from a bivariate normal population, are found in small samples. An approximation for large samples is given. The empirical results obtained at the Indian Statistical Institute compare favourably with the theoretical results.

Let Z be a normal random variable with mean m and variance σ^2 .
Let us denote the expectation of $|Z|$ by

$$E(|Z|) = F(m, \sigma^2). \quad \dots (1)$$

In case $m = 0$ we have
$$F(0, \sigma^2) = \sqrt{\frac{2}{\pi}} \sigma.$$

If σ^2 is a constant but m is a random variable with mean 0 and variance τ^2 then the expectation of $F(m, \sigma^2)$ can be expressed by

$$\begin{aligned} E(F(m, \sigma^2)) &= \frac{1}{2\pi\sigma\tau} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |x| e^{-\frac{1}{2} \left(\frac{(x-y)^2}{\sigma^2} + \frac{y^2}{\tau^2} \right)} dx dy \\ &= \sqrt{\frac{2}{\pi}} \sqrt{\sigma^2 + \tau^2} = \sqrt{\frac{2}{\pi}} \sigma \left(1 + F_0 \left(\frac{\tau}{\sigma} \right) \right) \end{aligned} \quad \dots (2)$$

where
$$F_0 \left(\frac{\tau}{\sigma} \right) = \sqrt{1 + \frac{\tau^2}{\sigma^2}} - 1. \quad \dots (3)$$

Now let Z_1, Z_2 be independent normal random variables with means m_1, m_2 respectively and with the same variance σ^2 .

Consider the statistic

$$a = \Delta(Z_1, Z_2) = \frac{|Z_1 Z_2|}{|Z_1| + |Z_2|} \quad \dots (4)$$

where

$$\Delta(a, b) = \begin{cases} 0 & \text{if } ab \geq 0 \\ 1 & \text{if } ab < 0 \end{cases}$$

Let us denote the expectation of a by

$$E(a) = G(m_1, m_2, \sigma^2). \quad \dots (5)$$

In case $m_1 = m_2 = 0$ we have $G(0, 0, \sigma^2) = \sqrt{\frac{2}{\pi}} \alpha_0 \sigma$

where

$$\begin{aligned} \alpha_0 &= \frac{1}{2\sqrt{2\pi}} \left\{ \int_{-\infty}^0 \int_0^{\infty} + \int_0^{\infty} \int_{-\infty}^0 \frac{|uv|}{|u|+|v|} e^{-(u^2+v^2)/2} du dv \right\} \\ &= \frac{1}{2} \left(1 - \frac{1}{\sqrt{2}} \log(1+\sqrt{2}) \right) = 0.1882. \quad \dots (6) \end{aligned}$$

If σ^2 is a constant but (m_1, m_2) is a two dimensional normal random variable with means $(0, 0)$, variances (τ_1^2, τ_2^2) and correlation coefficient ρ then the expectation of $G(m_1, m_2, \sigma^2)$ can be expressed by

$$\begin{aligned} E(G(m_1, m_2, \sigma^2)) &= \frac{1}{4\pi^2 \sigma^2 \tau_1^* \tau_2^* \sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^0 \int_0^{\infty} + \int_0^{\infty} \int_{-\infty}^0 \right\} \\ &\quad \times \frac{x|y|}{|x|+|y|} e^{-\frac{1}{2\sigma^2} \{(x-u)^2 + (y-v)^2\}} - \frac{1}{2(1-\rho^2)} \left\{ \frac{u^2}{\tau_1^2} - 2\frac{\rho uv}{\tau_1 \tau_2} + \frac{v^2}{\tau_2^2} \right\} dudv dx dy \\ &= \frac{1}{\pi \sqrt{\Delta}} \int_0^{\infty} \int_0^{\infty} \frac{xy}{x+y} e^{-\frac{1}{2\Delta} \{(\sigma^2 + \tau_2^2)x^2 + 2\rho\tau_1\tau_2 xy + (\sigma^2 + \tau_1^2)y^2\}} dx dy \end{aligned}$$

where

$$\Delta = (\sigma^2 + \tau_1^2)(\sigma^2 + \tau_2^2) - \rho^2 \tau_1^2 \tau_2^2.$$

By means of polar coordinates we have also

$$E(G(m_1, m_2, \sigma^2)) = \frac{\Delta}{\sqrt{2\pi}} \int_0^{\pi/2} \frac{\sin \theta \cos \theta}{\sin \theta + \cos \theta} \frac{d\theta}{f(\theta)^{3/2}}$$

where

$$f(\theta) = (\sigma^2 + \tau_2^2) \cos^2 \theta + 2\rho\tau_1\tau_2 \sin \theta \cos \theta + (\sigma^2 + \tau_1^2) \sin^2 \theta.$$

Let us put

$$\gamma = \sigma^2 + \frac{\tau_1^2 + \tau_2^2}{2}, \delta = \frac{\tau_2^2 - \tau_1^2}{2\sigma^2 + \tau_1^2 + \tau_2^2}, \epsilon = \frac{2\rho\tau_1\tau_2}{2\sigma^2 + \tau_1^2 + \tau_2^2}. \quad \dots (7)$$

If $\frac{\tau_i^2}{\sigma^2}$ ($i = 1, 2$) are small we can expand $f(\theta)^{-\frac{3}{2}} = \gamma^{-\frac{3}{2}} (1 + \epsilon \sin 2\theta + \delta \cos 2\theta)^{-\frac{3}{2}}$

ON THE EXPECTATION OF ERROR AREA IN FRACTILE GRAPHICAL ANALYSIS

in power series in ϵ and δ and then we have

$$E(G(m_1, m_2, \sigma^2)) = \sqrt{\frac{2}{\pi}} \frac{\Delta}{\gamma^{\frac{1}{2}}} \left(\alpha_0 - \frac{1-3\alpha_0}{2} \epsilon + \frac{15\alpha_0-1}{8} \epsilon^2 + \frac{1}{8} \delta^2 + \dots \right).$$

Substituting Δ and γ by $\Delta = \gamma^2(1-\epsilon^2-\delta^2)$,

$$\gamma^{\frac{1}{2}} = \left(1 + \frac{1}{2} F_0 \left(\frac{\tau_1}{\sigma} \right) + \frac{1}{2} F_0 \left(\frac{\tau_2}{\sigma} \right) \right) \times \left(1 - \frac{\delta^2}{8} + \dots \right) \text{ we have finally,}$$

$$E(G(m_1, m_2, \sigma^2)) = \sqrt{\frac{2}{\pi}} \alpha_0 \sigma \left(1 + \frac{1}{2} F_0 \left(\frac{\tau_1}{\sigma} \right) + \frac{1}{2} F_0 \left(\frac{\tau_2}{\sigma} \right) \right) \\ \times \left(1 - \frac{1-3\alpha_0}{2\alpha_0} \epsilon + \frac{7\alpha_0-1}{8\alpha_0} \epsilon^2 - \frac{9\alpha_0-1}{8\alpha_0} \delta^2 + \dots \right).$$

If we neglect the higher order terms in $\frac{\tau_i^2}{\sigma^2}$ ($i = 1, 2$) then we have

$$E(G(m_1, m_2, \sigma^2)) \doteq \sqrt{\frac{2}{\pi}} \alpha_0 \sigma \cdot \left(1 + \frac{1}{2} F_0 \left(\frac{\tau_1}{\sigma} \right) + \frac{1}{2} F_0 \left(\frac{\tau_2}{\sigma} \right) - \frac{1-3\alpha_0}{2\alpha_0} \epsilon \right). \quad \dots (8)$$

Let (x_i, y_i) $i = 1, \dots, n$; (x_i^*, y_i^*) $i = 1, \dots, n$ be $2n$ independent observations on the same two dimensional normal variable with means, variances and correlation coefficient $\mu, \nu, \sigma_1^2, \sigma_2^2$ and ρ . Rearrange the first sample as follows

$$(x_{(1)}, y_{(1)}), \dots (x_{(n)}, y_{(n)}) \text{ so that } x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

Let $n = mg$ where m and g are integers. Define

$$\bar{y}_i = \frac{1}{m} \sum_{r=(i-1)m+1}^{im} y_{(r)}; \quad i = 1, \dots, g$$

$$\bar{x}_i = \frac{1}{m} \sum_{r=(i-1)m+1}^{im} x_{(r)}; \quad i = 1, \dots, g,$$

$\bar{y}_i^*$ and $\bar{x}_i^*$ are defined similarly. Let us denote $\bar{y}_i - \bar{y}_i^*$ by z_i and $\frac{\rho\sigma_y}{\sigma_x} (\bar{x}_i - \bar{x}_i^*)$ by w_i

for $i = 1, \dots, g$.

Plotting the points $\bar{y}_1, \dots, \bar{y}_g$ at the equidistant points $1, \dots, g$ and joining them successively we get a curve G known as the fractile graph of the first sample. The fractile graph G^* of the second sample is likewise obtained. The area A between

the graphs G and G^* between the coordinates at 1 and g is defined as the error area between the two samples. We easily construct the algebraic expression for the error area A as

$$A = \sum_{i=1}^{g-1} \left\{ \frac{1}{2} |z_i| + \frac{1}{2} |z_{i+1}| - \Delta(z_i, z_{i+1}) \frac{|z_i z_{i+1}|}{|z_i| + |z_{i+1}|} \right\}. \quad \dots (9)$$

Let $(x_1, \dots, x_n, x_1^*, \dots, x_n^*)$ be fixed. Then we see that $z_1, \dots, z_g$ are independently normally distributed with the same variance $\frac{2\sigma_y^2(1-\rho^2)}{m}$ and means $w_1, \dots, w_g$.

Applying (1) and (5) we get the conditional expectation

$$E(A | x_1, \dots, x_n, x_1^*, \dots, x_n^*).$$

Hence the unconditional expectation is obtained as

$$E(A) = \sum_{i=1}^{g-1} \left\{ \frac{1}{2} E(F(w_i, \sigma^2)) + \frac{1}{2} E(F(w_{i+1}, \sigma^2)) - E(G(w_i, w_{i+1}, \sigma^2)) \right\}. \quad \dots (10)$$

Let us consider the case where m is large. Then we can consider $w_1, \dots, w_g$ as normal variates with means $(0, \dots, 0)$. Let variances and correlation coefficients of $w_1, \dots, w_g$ be denoted by τ_i^2 ($i = 1, \dots, g$) and ρ_{ij} ($i, j = 1, \dots, g$). Assuming $\frac{\tau_i^2}{\sigma^2}$ ($i = 1, \dots, g$) small we can apply (2) and (8), and we get

$$E(A) \doteq \frac{g-1}{\sqrt{m}} \frac{2}{\sqrt{\pi}} \sigma_y \sqrt{1-\rho^2} \left\{ (1-\alpha_0)(1+B) + \frac{1-3\alpha_0}{2} C \right\} \quad \dots (11)$$

where $\alpha_0 = 0.1882$ as before and

$$B = \frac{1}{g-1} \sum_{i=1}^g \alpha_i F_0 \left(\frac{\tau_i}{\sigma} \right), \alpha_1 = \alpha_g = \frac{1}{2}, \alpha_2 = \dots = \alpha_{g-1} = 1 \quad \dots (12)$$

$$C = \frac{1}{g-1} \sum_{i=1}^{g-1} \frac{2\rho_{i,i+1} \tau_i \tau_{i+1}}{2\sigma^2 + \tau_i^2 + \tau_{i+1}^2}.$$

By the well known results of F. Mosteller $(x_{(m+1)}, \dots, x_{(n-m-1)})$ is approximately normally distributed with means $\mu_x + \sigma_x u_r$, variances $\sigma_x^2 \frac{\lambda_r(1-\lambda_r)}{nf_0(u_r)^2}$ and covariances $\sigma_x^2 \frac{\lambda_r(1-\lambda_s)}{nf_0(u_r)f_0(u_s)}$ ($r < s$)

where
$$\lambda_r = \frac{r}{n} \int_{-\infty}^{u_r} f_0(x) dx = \lambda_r, f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

ON THE EXPECTATION OF ERROR AREA IN FRACTILE GRAPHICAL ANALYSIS

Then we have

$$\tau_i^2 = \frac{2\sigma_y^2 \rho^2}{n} \quad k_i^2, \tau^2 = \frac{k_i^2 \rho^2}{g(1-\rho^2)}$$

$$k_i^2 = \frac{1}{m^2} \left\{ \sum_{r=(i-1)m+1}^{im} \frac{\lambda_r(1-\lambda_r)}{f_0(u_r)^2} + 2 \sum_{r < s} \frac{\lambda_r(1-\lambda_s)}{f_0(u_r) f_0(u_s)} \right\}$$

$$\rho_{i,i+1} k_i k_{i+1} = \left\{ \frac{1}{m} \sum_{r=(i-1)m+1}^{im} \frac{\lambda_r}{f_0(u_r)} \right\} \left\{ \frac{1}{m} \sum_{s=(im+1)}^{(i+1)m} \frac{1-\lambda_s}{f_0(u_s)} \right\}$$

and

$$F_0 \left(\frac{\tau_i}{\sigma} \right) = \sqrt{1 + \frac{k_i^2 \rho^2}{g(1-\rho^2)}} - 1 \quad \text{for } i = 2, \dots, g-1$$

$$\frac{2\rho_{i,i+1} \tau_i \tau_{i+1}}{2\sigma^2 + (\tau_i^2 + \tau_{i+1}^2)} = \frac{2\rho_{i,i+1} k_i k_{i+1} \rho^2}{2g(1-\rho^2) + \rho^2(k_i^2 + k_{i+1}^2)} \quad \text{for } i = 2, \dots, g-2. \quad \dots (13)$$

Moreover, if g is also sufficiently large $\frac{\tau_i^2}{\sigma^2}$ ($i = 1, \dots, g$) are certainly small and we can apply (8) as we have assumed before. We may also approximate the values k_i and $\rho_{i,i+1}$ by

$$k_i^2 = \frac{\lambda^{(i)}(1-\lambda^{(i)})}{f_0(u^{(i)})^2}, \quad \lambda^{(i)} = \frac{i}{g} - \frac{1}{2g} \int_0^{u^{(i)}} f_0(x) dx = \lambda^{(i)}$$

$$\rho_{i,i+1} k_i k_{i+1} = \frac{\lambda^{(i)}}{f_0(u^{(i)})} \frac{(1-\lambda^{(i+1)})}{f_0(u^{(i+1)})}. \quad \dots (14)$$

Though it is not correct we may apply these formulae also for $k_1, k_g, \rho_{1,2}$ and $\rho_{g-1,g}$ in B and C without effecting so much in the result since g is sufficiently large.

By (11), (12), (13) and (14) we can compute the value of $E(A)$.

Consider the empirical results obtained at the Indian Statistical Institute, (Mahalanobis, 1958). Samples were taken from a bivariate normal population with variances and correlation coefficient equal to 1, $\sqrt{2}$ and $\frac{1}{\sqrt{2}}$ respectively. The empirical result shows e.g.,

$$\begin{aligned} \frac{\sqrt{m}}{g-1} E(A) &= 1.09 \pm 0.04 \quad \text{for } g = 16, m = 12 \\ &= 1.07 \pm 0.05 \quad \text{for } g = 16, m = 24. \end{aligned}$$

The agreement between the figures is fairly good.

For the case $g = 16$ we find, from (11)–(14), that $B = 0.068$, $C = 0.092$ and $\frac{\sqrt{m}}{g-1}$

$$E(A) = 1.00.$$

It is suggested that a modified statistic namely $\frac{1}{2}|z_1| + |z_2| + \dots + |z_{g-1}| + \frac{1}{2}|z_g|$ may be used instead of the area. The terms involving the function G will be eliminated in the analysis.

REFERENCES

- MAHALANOBIS, P. C. (1958) : Lectures in Japan. Fractile Graphical Analysis : Indian Statistical Institute.
- MOSTELLER, F. (1945) : On some useful "inefficient" statistics. *Ann. Math. Stat.* **17**, 377-408.
- TAKEUCHI, K. (1961) : On some properties of the error area in fractile graph method. *Sankhyā*, Series A, **23**, 65-78.

Paper received : June, 1959.

Revised : December, 1960.

MINIMAX TESTS FOR THE RATE OF A POISSON PROCESS AND THE BIAS RATE OF A NORMAL PROCESS*

By J. V. BREAKWELL

Lockheed Aircraft Corporation, California, USA

SUMMARY. The problem of determining all the Bayes' procedures for distinguishing between two Poisson process rates, when the losses due to accepting the wrong rate are known and when the cost of observation is proportional to the observation time, is given an explicit solution. The selection of the minimax procedure is explained. The problem of distinguishing between two normal processes, with the same diffusion rate but different bias rates, is treated as a limiting case of the Poisson problem. The minimax solutions are also given for the extended problems in which a Poisson rate, or a normal process bias rate, can have any value. It is desired to decide whether this rate is above or below some critical rate, the loss due to a wrong decision being proportional to the difference between the actual rate and the critical rate. In some Poisson situations the minimax procedure is a mixed strategy.

1. INTRODUCTION

The author's interest in the Poisson and normal processes was first aroused by the possibility of using them in approximate descriptions of the sequential testing of a binomial parameter. The Poisson process is to be used when this parameter is small. They have, of course, a great number of other applications. It seems worthwhile, therefore, to solve a basic statistical-economic problem connected with them, namely, the determination of the minimax procedure for distinguishing between two process rates when the cost of observation is proportional to the time of observation of the process, and when the losses due to accepting the wrong rate are known. These solutions are described in Sections 2, 3, and 4. In the case of the normal process (Wiener process), the diffusion rate is assumed known, while the bias rate is in question. This paper also contains (Sections 5 and 6) the solutions of the important extensions of the above problems in which the process rate can now have any value (it being desired to decide whether this rate is above or below some critical rate) and in which the loss due to a wrong decision is proportional to the difference between the actual rate and the critical rate. The constants of proportionality in the loss function above and below the critical rate are not assumed to be the same. This form of loss function was originated by Wald (1950, p.9). This lack of economic symmetry allows for introduction of the right amount of bias in the test procedure when one type of wrong decision is economically more serious than the other. The solution of the extended problem for a Wiener process was actually obtained and used by the author sometime before solving the simpler dichotomous problem (Breakwell, 1954). This solution, in the case of economic symmetry, has been obtained independently by Moriguti (1955) and by Morris H. De Groot in a doctoral dissertation.

*A preliminary account of most of the results in this paper was given in a contributed paper entitled "Minimax Tests for the Parameter of a Poisson Process," presented at the Western Regional Meeting of the I.M.S. in Berkeley, July 1955.

The solution of the dichotomous problem for the Poisson process is a rather straightforward application, adapted to continuous observation, of the method presented in Chapter 10 of *Theory of Games and Statistical Decisions* by Girshick and Blackwell (1954). The method first obtains (Section 2) a Bayes' family of Sequential-Probability-Ratio tests (SPR tests) corresponding to various *a priori* weightings between the two possible Poisson rates, and then selects (Section 3) the minimax procedure by looking at the economically least favourable *a priori* weighting. The determination of the SPR tests, based on the notion of *a posteriori* risk (Wald, 1950; Girshick and Blackwell, 1954) requires first the solution of a problem about a "Poisson random walk with absorbing barriers" (see Girshick and Blackwell, 1954, p. 269) to obtain the operating characteristic (OC) and average observation time (AOT) for a general SPR test. These have already been obtained by Burman (1946) and by Dvoretzky, Kiefer and Wolfowitz (1953).

The solution of the dichotomous problem for the normal process may be obtained by the same methods. The solution of the corresponding random walk problem is considerably easier (using methods used by Wald (1947)) and is given by Dvoretzky, Kiefer and Wolfowitz (1953) in which the diffusion rate is taken as unity. Their formula for ET, however, needs clarification. The subsequent determination of the Bayes' family of SPR tests in terms of the economic losses is, however, more tedious than in the Poisson problem. The author of this paper has preferred, therefore, to treat the normal process somewhat heuristically as a limiting case of the Poisson process and to derive the solution of the dichotomous normal problem from that of the Poisson problem.

The minimax solution of the dichotomous Poisson problem, unlike that of the normal problem, can lead to a mixed test procedure, and this carries over to a certain economic region for the extended problem.

The minimax solution of the extended Poisson problem has been applied by the author to the problem of testing for a low fraction of defectives (1956); except that, in economic situations where the minimax strategy is mixed, the author preferred to recommend a "minimax pure strategy," selected from the set of all possible SPR tests. This set was augmented by the possible outright rejection of the hypothesis that the Poisson rate is below critical.

2. THE BAYES' PROBLEM FOR DISTINGUISHING TWO POISSON PROCESSES

First of all, we may describe a Poisson process with rate λ by saying that a certain occurrence has probability $\lambda\Delta t + o(\Delta t)$ of occurring during any small time interval Δt , regardless of how many times it has previously occurred. The number N of occurrences during any time interval thus satisfies the Poisson formula :

$$\Pr \{N, t\} = \frac{(\lambda t)^N}{N!} e^{-\lambda t} \quad \dots \quad (2.1)$$

We wish to observe this process and to test the hypothesis $H_1 : \lambda = \lambda_1$ against the alternative hypothesis $H_2 : \lambda = \lambda_2 > \lambda_1$. Let us assume that the loss due to accepting

MINIMAX TESTS FOR POISSON PROCESS AND NORMAL PROCESS BIAS RATES

H_2 when H_1 is true is a known positive quantity W_{12} ; and that the loss due to accepting H_1 when H_2 is true is another known positive quantity W_{21} ; and finally, that the cost of observation is c per unit of time.

If we are given an *a priori* weighting $\eta : 1 - \eta$ of $H_1 : H_2$, the Bayes' procedure, say Δ_η , is characterized among all test procedures Δ as the one which will minimize the expected risk :

$$r(\eta, \Delta) = \eta r_1 + (1 - \eta) r_2, \quad \dots \quad (2.2)$$

where

$$\begin{aligned} r_1 &= W_{12}P(2|1) + ci(1) \\ r_2 &= W_{21}P(1|2) + ci(2), \end{aligned} \quad \dots \quad (2.3)$$

and where $P(j|i)$ denotes the probability of accepting H_j when H_i is true, while $i(i)$ denotes the average observation time (AOT) when H_i is true.

Now it is known generally [Blackwell and Girshick (1954); Wald (1950)] that the Bayes' procedure Δ_η may be characterized as follows :

Starting with *a priori* weighting η in favour of H_1 , compute continually the *a posteriori* weighting $\eta(t)$ after time t , and

- accept H_1 if $\eta(t) \geq$ a certain value η_1 greater than zero and less than one;
 - accept H_2 if $\eta(t) \leq$ a certain value η_2 greater than zero and not exceeding η_1 ;
 - continue to observe if $\eta_2 < \eta(t) < \eta_1$.
- ... (2.4)

In particular, if the original $\eta > \eta_1$, this procedure requires us to accept H_1 outright, while if $\eta < \eta_2$ we are to accept H_2 outright. The equality signs in (2.4) are optional.

Now, because of the convexity of the risk $r(\eta, \Delta_\eta)$ as a function of η , it follows that η_1 and η_2 are determined uniquely by λ_1, λ_2 and the economic quantities W_{12}, W_{21}, c . The actual determination will be described later. In particular, it may happen that $\eta_2 = \eta_1$ and that the Bayes' procedure is always an outright decision.

Next, the *a posteriori* weighting $\eta(t)$ in favour of H_1 is expressible in the form

$$\eta(t) = \frac{\eta}{\eta + (1 - \eta)L(t)} \quad \dots \quad (2.5)$$

where $L(t)$ is the likelihood ratio after time t for distinguishing hypothesis H_2 from hypothesis H_1 . The Bayes' procedure Δ_η , when $\eta_2 < \eta < \eta_1$, is thus an SPR test describable as follows :

- accept H_1 if and when $L(t) \leq \frac{\eta(1 - \eta_1)}{(1 - \eta)\eta_1}$
 - accept H_2 if and when $L(t) > \frac{\eta(1 - \eta_2)}{(1 - \eta)\eta_2}$
 - continue to observe as long as $\frac{\eta(1 - \eta_1)}{(1 - \eta)\eta_1} < L(t) \leq \frac{\eta(1 - \eta_2)}{(1 - \eta)\eta_2}$.
- ... (2.6)

The unsymmetrical placing of the equality signs in (2.6) is optional in the general characterization but is adopted here for convenience in what follows.

Next, the likelihood ratio $L(t)$ for distinguishing the Poisson process rate λ_2 from the rate λ_1 is :

$$L(t) = \left\{ \frac{(\lambda_2 t)^N}{N!} e^{-\lambda_2 t} \right\} \div \left\{ \frac{(\lambda_1 t)^N}{N!} e^{-\lambda_1 t} \right\} \quad \dots (2.7)$$

It follows that the SPR test may now be characterized as follows :
(See Figure 1.)

$$\left. \begin{array}{l} \text{accept } H_1 \text{ if and when } y(t) \geq z; \\ \text{accept } H_2 \text{ if and when } y(t) < 0; \\ \text{continue to observe as long as } 0 \leq y(t) < z; \\ \text{where } y(t) = y + \mu t - N, \end{array} \right\} \quad \dots (2.8)$$

$$\left. \begin{array}{l} \text{and where } \mu = \frac{\lambda_2 - \lambda_1}{\ln(\lambda_2/\lambda_1)}, \\ y = \ln \frac{\eta(1-\eta_2)}{(1-\eta)\eta_2} \div \ln\left(\frac{\lambda_2}{\lambda_1}\right), \\ z = \ln \frac{\eta_1(1-\eta_2)}{(1-\eta_1)\eta_2} \div \ln\left(\frac{\lambda_2}{\lambda_1}\right) \end{array} \right\} \quad \dots (2.9)$$

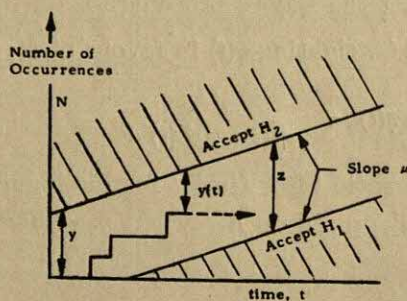


Figure 1.

The determination of η_1 and η_2 in terms of λ_i , W_{ij} , c , will be based on the facts (which follows from the continuity of the Bayes' risk $r(\eta, \Delta_\eta)$ as a function of η) that whenever $\eta(t) = \eta_1$, the statistician does equally well by accepting H_1 outright or by proceeding with the appropriate SPR test, and that whenever $\eta(t) = \eta_2$ the statistician does equally well by accepting H_2 outright or by proceeding with the appropriate SPR test. In other words, the placing of the equality signs in (2.4) and (2.6) is not crucial,

MINIMAX TESTS FOR POISSON PROCESS AND NORMAL PROCESS BIAS RATES

Now, it is apparent from Figure 1 that during any infinitesimal time interval Δt , the distance $y(t)$ from the upper boundary will either increase by an amount $\mu\Delta t$ (this with probability $1 - \lambda_i\Delta t + o(\Delta t)$) or will decrease by $1 + o(\Delta t)$ (this with probability $\lambda_i\Delta t$) where $i = 1$ or 2 , according as H_1 or H_2 is true. It follows that the expected further risk, if $\eta(t) = \eta_1$ and if the statistician proceeds with the appropriate SPR test so that $y = z$, is

$$c\Delta t + \eta_1\{(1 - \lambda_1\Delta t)0 + \lambda_1\Delta t[W_{12}P(2|1, z-1) + c\bar{i}(1, z-1)]\} + \\ + (1 - \eta_1)\{(1 - \lambda_2\Delta t)W_{21} + \lambda_2\Delta t[W_{21}P(1|2, z-1) + c\bar{i}(2, z-1)]\} + o(\Delta t),$$

where we have written $P(j|i, y)$ and $\bar{i}(i, y)$ to indicate the dependence of $P(j|i)$ and $\bar{i}(i)$ on y , but have not indicated their dependence on z . Equating this to the expected risk $(1 - \eta_1)W_{21}$ [Blackwell and Girshick, (1954), p. 269] corresponding to outright acceptance of H_1 , dividing by Δt , and letting $\Delta t \rightarrow 0$, we obtain

$$c + \eta_1\lambda_1[W_{12}P(2|1, z-1) + c\bar{i}(1, z-1)] \\ = (1 - \eta_1)\lambda_2[W_{21}P(2|2, z-1) + c\bar{i}(2, z-1)]. \quad \dots (2.10)$$

Furthermore, the expected further risk, if $\eta(t) = \eta_2$ and if the statistician proceeds with the appropriate SPR test so that $y = 0$, is

$$c\Delta t + \eta_2\{(1 - \lambda_1\Delta t)[W_{12}P(2|1, \mu\Delta t) + c\bar{i}(1, \mu\Delta t)] + \lambda_1\Delta t.W_{12}\} + \\ + (1 - \eta_2)\{(1 - \lambda_2\Delta t)[W_{21}P(1|2, \mu\Delta t) + c\bar{i}(2, \mu\Delta t)] + \lambda_2\Delta t.0\} + o(\Delta t).$$

Equating this to the expected risk $\eta_2 W_{12}$, corresponding to outright acceptance of H_2 , and letting $\Delta t \rightarrow 0$, we obtain

$$\eta_2[W_{12}P(1|1, 0) + c\bar{i}(1, 0)] = (1 - \eta_2)[W_{21}P(1|2, 0) + c\bar{i}(2, 0)]. \quad \dots (2.11)$$

Note that according to our convention on the equalities in (2.6) and (2.8), the SPR test starting with $y = 0$ is not trivial, being quite different from outright acceptance of H_2 .

The OC function is known from Burman (1946) and Dvoretzky, Kiefer and Wolfowitz (1953). Thus :

$$1 - P(2|i, y) = P(1|i, y) = \frac{F\left(y, \frac{\lambda_i}{\mu}\right)}{F\left(z, \frac{\lambda_i}{\mu}\right)}, \quad \dots (2.12)$$

where

$$F(y, x) = e^{xy}e^*(-xe^{-x}, y), \quad \dots (2.13)$$

and $e^*(a, y)$ denotes the "Tapered Exponential Function" [see Hammersley (1953)] defined by

$$e^*(a, y) = \sum_{k=0}^{[y]} \frac{1}{k!} \{a(y-k)\}^k, \quad \dots (2.14)$$

$[y]$ denoting the largest integer not exceeding y .

The AOT function is also obtained in Burman (1946) and Dvoretzky, Kiefer and Wolfowitz (1953). It is :

$$i(i, y) = \frac{1}{\lambda_i} \left\{ \frac{F\left(y, \frac{\lambda_i}{\mu}\right)}{F\left(z, \frac{\lambda_i}{\mu}\right)} G\left(z, \frac{\lambda_i}{\mu}\right) - G\left(y, \frac{\lambda_i}{\mu}\right) \right\}, \quad \dots (2.15)$$

where
$$G(y, x) = \sum_{i=0}^{[y]} \{F(y-i, x) - 1\}. \quad \dots (2.16)$$

Insertion of (2.12) and (2.15) into (2.10) and (2.11), taking into account that $G\left(z, \frac{\lambda_i}{\mu}\right)$ and $F\left(z, \frac{\lambda_i}{\mu}\right)$ [$i = 1, 2$] are increasing functions of z as is $F\left(z, \frac{\lambda_2}{\mu}\right) - \left(1 - \frac{\lambda_1}{\lambda_2}\right) G\left(z, \frac{\lambda_2}{\mu}\right)$, leads, after some manipulation, to the main result of this Section :

If
$$\left\{ \frac{(\lambda_2 - \lambda_1)W_{21}}{c} - 1 \right\} \left\{ \frac{(\lambda_2 - \lambda_1)W_{12}}{c} - 1 \right\} > 1,$$

then : (a) There are unique numbers z , η_1 , and $\eta_2 (< \eta_1)$ obtainable (after first solving (2.9) for μ) from the following three equations :

$$\begin{aligned} & \left\{ \frac{(\lambda_2 - \lambda_1)W_{21}}{c} + \left(1 - \frac{\lambda_1}{\lambda_2}\right) G\left(z, \frac{\lambda_2}{\mu}\right) - F\left(z, \frac{\lambda_2}{\mu}\right) \right\} \\ & \left\{ \frac{(\lambda_2 - \lambda_1)W_{12}}{c} - \left(\frac{\lambda_2}{\lambda_1} - 1\right) G\left(z, \frac{\lambda_1}{\mu}\right) - F\left(z, \frac{\lambda_1}{\mu}\right) \right\} \quad \dots (2.17) \\ & = F\left(z, \frac{\lambda_2}{\mu}\right) F\left(z, \frac{\lambda_1}{\mu}\right) \end{aligned}$$

subject to the restriction that both brackets { } of the left member be positive;

$$\frac{\eta_1}{1 - \eta_1} = \frac{\lambda_2 W_{21} + c G\left(z - 1, \frac{\lambda_2}{\mu}\right) - c}{\lambda_1 W_{12} - c G\left(z - 1, \frac{\lambda_1}{\mu}\right) + c} \quad \dots (2.18)$$

$$\frac{1 - \eta_2}{\eta_2} = \left(\frac{\lambda_2}{\lambda_1}\right)^{z+1} \frac{\lambda_1 W_{12} - c G\left(z, \frac{\lambda_1}{\mu}\right)}{\lambda_2 W_{21} + c G\left(z, \frac{\lambda_2}{\mu}\right)}. \quad \dots (2.19)$$

(b) If $\eta_2 < \eta < \eta_1$, the Bayes' procedure $\Delta = \Delta_\eta$ for distinguishing between rates λ_1 and $\lambda_2 (> \lambda_1)$, given an *a priori* weighting $\eta : 1 - \eta$ in favour of λ_1 , is the SPR test defined by the parameters μ , z , and y . Here z is obtained from (a) and y from (2.9). For values of η outside the interval (η_2, η_1) , of course, the Bayes' procedure Δ_η is an outright decision according to (2.4).

MINIMAX TESTS FOR POISSON PROCESS AND NORMAL PROCESS BIAS RATES

The solution of the transcendental equation (2.17) for z may be facilitated by a short table of the tapered exponential function given in Hammersley (1953).

If, on the other hand, $\left\{ \frac{(\lambda_2 - \lambda_1)W_{21}}{c} - 1 \right\} \left\{ \frac{(\lambda_2 - \lambda_1)W_{12}}{c} - 1 \right\} \leq 1$, an outright decision is always preferable, according to the scheme:

accept H_1 when $\eta < \eta_0$; accept H_2 when $\eta > \eta_0$;

where η_0 is given by : $\eta_0 W_{12} = (1 - \eta_0) W_{21}$, ... (2.20)

the correct decision being optional when $\eta = \eta_0$.

It should be noted that if

$$\left\{ \frac{(\lambda_2 - \lambda_1)W_{21}}{c} - 1 \right\} \left\{ \frac{(\lambda_2 - \lambda_1)W_{12}}{c} - 1 \right\} > 1,$$

but
$$\left(\frac{\lambda_2 W_{21}}{c} - 1 \right) \left(\frac{\lambda_1 W_{12}}{c} + 1 \right) \leq \frac{\lambda_1}{\lambda_2 - \lambda_1} \left(\frac{\lambda_2}{\lambda_1} \right)^{\frac{\lambda_2}{\lambda_2 - \lambda_1}} \left(\frac{\lambda_1 W_{12} + \lambda_2 W_{21}}{c} \right),$$

then the parameter z lies in the interval : $0 < z \leq 1$, and the SPR test becomes a "limited sequential test", more simply described in terms of a single parameter :

$$T = \frac{z - y}{\mu}, \quad \dots (2.21)$$

as follows :

test for a time t not exceeding T ;

accept H_2 if and when anything occurs ;

accept H_1 otherwise.

The acceptance probabilities and average observation time are simply :

$$1 - P(2|i) = P(1|i) = e^{-\lambda_i T}, \quad \dots (2.22)$$

and

$$i(i) = \frac{1}{\lambda_i} (1 - e^{-\lambda_i T}). \quad \dots (2.23)$$

In place of the test parameter y , which increased from 0 to z as η increased from η_2 to η_1 , we now have the test parameter T , which decreases from T_2 to 0 where $T_2 = \frac{z}{\mu}$ and is given by the transcendental equation :

$$\left\{ W_{12} - \frac{c}{\lambda_1} (e^{\lambda_1 T_2} - 1) \right\} (\lambda_2 W_{21} - c) = (\lambda_1 W_{12} + c) \left\{ W_{21} + \frac{c}{\lambda_2} (e^{\lambda_2 T_2} - 1) \right\}. \quad \dots (2.24)$$

η_1 is now obtained from :

$$\frac{\eta_1}{1-\eta_1} = \frac{\lambda_2 W_{21} - c}{\lambda_1 W_{12} + c}, \quad \dots (2.25)$$

while η_2 is obtained from :

$$\frac{1-\eta_2}{\eta_2} = \frac{W_{12} e^{-\lambda_1 T_2} - \frac{c}{\lambda_1} (1 - e^{-\lambda_1 T_2})}{W_{21} e^{-\lambda_2 T_2} + \frac{c}{\lambda_2} (1 - e^{-\lambda_2 T_2})}. \quad \dots (2.26)$$

Finally, the test parameter T is determined as a function of η over the interval (η_2, η_1) by means of :

$$T = \frac{1}{\lambda_2 - \lambda_1} \ln \frac{\eta_1(1-\eta)}{(1-\eta_1)\eta}. \quad \dots (2.27)$$

Formulae (2.21) through (2.27) are applicable to the case in which $\lambda_1 = 0$, provided that the right members of (2.23) and (2.26) and the left member of (2.24) are allowed to assume their obvious limiting forms. Indeed, when $\lambda_1 = 0$, if an SPR test is to be a Bayes' procedure, it must necessarily be a "limited sequential test" as may easily be demonstrated by a separate treatment of the likelihood ratio $L(t)$.

3. THE MINIMAX PROCEDURE FOR DISTINGUISHING TWO POISSON PROCESSES

Let us suppose first that $\left\{ \frac{(\lambda_2 - \lambda_1)W_{21}}{c} - 1 \right\} \left\{ \frac{(\lambda_2 - \lambda_1)W_{12}}{c} - 1 \right\} > 1$. Then the minimized Bayes' risk $r(\eta, \Delta_\eta) = \min r(\eta, \Delta)$, obtained in Section 2 over the η -interval (η_2, η_1) , has a slope given by :

$$\frac{d}{d\eta} r(\eta, \Delta_\eta) = r_1 - r_2. \quad \dots (3.1)$$

This follows from the fact that $r(\eta, \Delta)$, restricted to test procedures Δ which are SPR tests, must be stationary at Δ_η with respect to the test parameters μ, z and y , only the last of which is a function of η .

Furthermore, the difference $r_2 - r_1$ is a monotonically decreasing function of η .

For values of η outside the interval (η_2, η_1) we have :

$$\text{for } \eta > \eta_1, r(\eta, \Delta_\eta) = (1-\eta)W_{21}; \text{ for } \eta < \eta_2, r(\eta, \Delta_\eta) = \eta W_{12}. \quad \dots (3.2)$$

But $y(\eta_1+0) = z$ and $r_1(\mu, z, z) - r_2(\mu, z, z) = -W_{21}$. On the other hand, $y(\eta_2+0) = 0$, but $r_1(\mu, z, 0) - r_2(\mu, z, 0) < W_{12}$. Thus the graph of $r(\eta, \Delta_\eta)$ versus η over the whole interval $(0,1)$ has everywhere a non-increasing slope which is discontinuous at $\eta = \eta_2$.

This graph thus takes one of the general forms in Figures 2 and 3 according as $d/d\eta r(\eta, \Delta_\eta)$, evaluated at $\eta = \eta_2+0$, is or is not non-negative. Because of (3.1)

MINIMAX TESTS FOR POISSON PROCESS AND NORMAL PROCESS BIAS RATES
 we see that Figure 2 is typical if $r_1(\mu, z, 0) \geq r_2(\mu, z, 0)$; otherwise Figure 3 is typical.

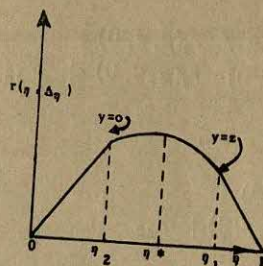


Figure 2.

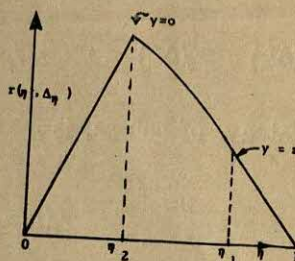


Figure 3.

The SPR tests described in Section 2 have been selected in the light not only of assumed known economic losses W_{12} and W_{21} and an assumed known test cost c per unit time, but also of an assumed known *a priori* weighting $\eta : 1 - \eta$ between a Poisson rate λ_1 and a rate λ_2 .

If, instead, η is not known, a "safe" course is to look for the minimax procedure (relative to the two rates λ_1 and λ_2), i.e., a procedure $\Delta = \Delta^*$ which will minimize $\max(r_1(\Delta), r_2(\Delta))$.

In situations typified by Figure 2, the minimax test procedure is simply the SPR test which corresponds to the abscissa η^* of the highest point on the graph, which test is defined by a test parameter $y = y^*$ (in addition to the test parameters μ and z previously defined) given by the further transcendental equation :

$$r_1(\mu, z, y^*) = r_2(\mu, z, y^*) = r^*, \text{ say.} \quad \dots (3.3)$$

Indeed, this SPR test Δ^* compares favourably with any other conceivable test procedure Δ , in the sense that

$$\eta^* r_1(\Delta) + (1 - \eta^*) r_2(\Delta) \geq \eta^* r_1(\Delta^*) + (1 - \eta^*) r_2(\Delta^*),$$

from which it follows, because of (3.3), that $\max(r_1(\Delta), r_2(\Delta)) \geq r^*$.

In those situations typified by Figure 3, the minimax procedure is a mixed procedure Δ_m^* defined as follows : either accept H_2 outright, this with weight ξ^* , or apply the SPR test defined by $y = 0$ (and by the previously defined μ and z), this with weight $1 - \xi^*$, where ξ^* is given by :

$$\xi^* = \frac{r_2(\mu, z, 0) - r_1(\mu, z, 0)}{r_2(\mu, z, 0) - r_1(\mu, z, 0) + W_{12}} \quad \dots (3.4)$$

which is the statistician's solution in the following 2×2 game between the statistician and "Nature":

		Nature		
		η	$1 - \eta$	
statistician	ξ	accept H_2 outright	W_{12}	0
	$1 - \xi$	SPR test with $y = 0$	$r_1(\mu, z, 0)$	$r_2(\mu, z, 0)$

... (3.5)

The mixed strategy Δ_m^* leads, as is easily shown to a risk independent of Nature's choice of Process rate λ_i , namely :

$$r_1(\Delta_m^*) = r_2(\Delta_m^*) = r^*, \text{ say, } = \frac{W_{12}r_2(\mu, z, 0)}{r_2(\mu, z, 0) - r_1(\mu, z, 0) + W_{12}} \quad \dots (3.6)$$

Nature's best strategy, of course, is

$$\eta = \eta_2 = \frac{r_2(\mu, z, 0)}{r_2(\mu, z, 0) - r_1(\mu, z, 0) + W_{12}} \quad \dots (3.7)$$

in agreement with (2.11), which leads to a risk equal to r^* in (3.6), whether the statistician accepts H_2 outright or uses the SPR test with $y = 0$.

We verify that the above mixed strategy is minimax by observing that no other strategy Δ , mixed or otherwise, can lead to a value for $\eta_2 r_1(\Delta) + (1 - \eta_2) r_2(\Delta)$ lower than $\eta_2 r_1(\Delta_m^*) + (1 - \eta_2) r_2(\Delta_m^*)$.

From this, because of (3.6) it follows again that

$$\max(r_1(\Delta), r_2(\Delta)) \geq r^*.$$

Suppose, finally, that

$$\left\{ \frac{(\lambda_2 - \lambda_1)W_{21}}{c} - 1 \right\} \left\{ \frac{(\lambda_2 - \lambda_1)W_{12}}{c} - 1 \right\} \leq 1.$$

Then, as pointed out at the end of Section 2, the Bayes' procedure is always an outright decision. The graph of $r(\eta, \Delta_\eta)$ vs η takes the simple form shown in Figure 4.

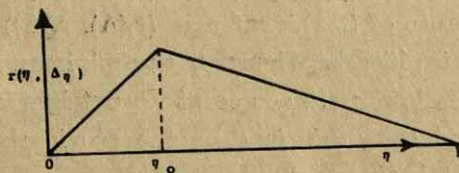


Figure 4.

The minimax strategy Δ_m^* is here the solution of the 2×2 game,

		Nature		
		η	$1 - \eta$	
statistician	ξ	accept H_2 outright	W_{12}	0
	$1 - \xi$	accept H_1 outright	0	W_{21}

... (3.8)

MINIMAX TESTS FOR POISSON PROCESS AND NORMAL PROCESS BIAS RATES

Namely :

$$\xi^* = \frac{W_{21}}{W_{21} + W_{12}} \quad \dots (3.9)$$

which again leads to a risk independent of nature's choice :

$$r^* = r_1(\Delta_m^*) = r_2(\Delta_m^*) = \frac{W_{12}W_{21}}{W_{21} + W_{12}} \quad \dots (3.10)$$

Nature's best strategy, of course, is $\eta = \eta_0$ given by (2.20), which again leads to a risk equal to r^* in (3.10), independent of the statistician's choice.

4. THE NORMAL PROCESS PROBLEM AS A LIMITING CASE

If y and $z - y$ are large and λ_1 not too small, the decision to accept H_1 or the decision to accept H_2 will take place only after a large number of occurrences. If t is a typical time of decision, the quantities $\lambda_1 t$ and $\lambda_2 t$ are thus large, so that the Poisson distribution of the number of occurrences in time t may be approximated by a normal distribution with mean and variance $\lambda_i t$ ($i = 1, 2$). If, in addition, the difference $|\lambda_2 - \lambda_1| \ll \lambda_1, \lambda_2$, it follows not only that the random walk in Figure 1 may be regarded as a "normal diffusion process" or "fundamental diffusion process," with a bias rate, [Chandrasekhar (1943), Uhlenbeck (1945)], but also that the two parameters λ_1 and λ_2 correspond to essentially the same diffusion rate. In other words, the problem in distinguishing two Poisson processes approaches the problem of distinguishing between two "Wiener processes" (Dvoretzky, Kiefer and Wolfowitz, 1953) with different bias rates but the same diffusion rate. To see this in greater detail, let the rates λ_1 and λ_2 be expressed in terms of two rates R and Δ_m and a number γ (which will later tend to ∞) as follows :

$$\lambda_1 + \lambda_2 = 2\gamma R, \quad \lambda_2 - \lambda_1 = \Delta_m \sqrt{\gamma}. \quad \dots (4.1)$$

Furthermore, let λ_0 be another rate differing from λ_1 or λ_2 by a quantity of order $\sqrt{\gamma}$, and let

$$m_i = \frac{\lambda_i - \lambda_0}{\sqrt{\gamma}} \quad \dots (4.2)$$

so that (because of (4.1))

$$m_2 - m_1 = \Delta_m, \quad \dots (4.3)$$

as implied by the notation. Then, as $\gamma \rightarrow \infty$, the distribution of the random quantity $X(t)$, defined by

$$X(t) = \frac{N - \lambda_0 t}{\sqrt{\gamma}} \quad \dots (4.4)$$

tends toward a normal distribution with variance Rt and mean $m_1 t$ or $m_2 t$, according as H_1 or H_2 is true. In other words, the time series $X(t)$ is asymptotically a normal (Wiener) process with variance R per unit time and bias m_1 or m_2 per unit time, according as H_1 or H_2 is true.

Meanwhile, from (4.1) and the first equation (2.9), it follows that

$$\left. \begin{aligned} \mu &= R\gamma \left\{ 1 + o\left(\frac{1}{\gamma}\right) \right\} = \frac{\lambda_1 + \lambda_2}{2} \left\{ 1 + o\left(\frac{1}{\gamma}\right) \right\}, \\ \frac{\lambda_1}{\mu} &= 1 - \frac{\Delta m}{2R\sqrt{\gamma}} + o\left(\frac{1}{\gamma}\right), \\ \frac{\lambda_2}{\mu} &= 1 + \frac{\Delta m}{2R\sqrt{\gamma}} + o\left(\frac{1}{\gamma}\right). \end{aligned} \right\} \quad \dots (4.5)$$

Let us next introduce test parameters Y and Z , defined by

$$\left. \begin{aligned} Y &= y/\sqrt{\gamma}, \\ Z &= z/\sqrt{\gamma} \end{aligned} \right\} \quad \dots (4.6)$$

By means of (4.2), (4.4), (4.6) and the first equation (4.5), the SPR test characterized in (2.8) may be re-characterized (as $\gamma \rightarrow \infty$) as follows : (Figure 5) :

$$\text{Accept } H_1 \text{ if and when } X(t) \leq m_T t - (Z - Y); \quad \dots (4.7)$$

$$\text{Accept } H_2 \text{ if and when } X(t) > m_T t + Y;$$

where

$$m_T = \frac{m_1 + m_2}{2} \quad \dots (4.8)$$

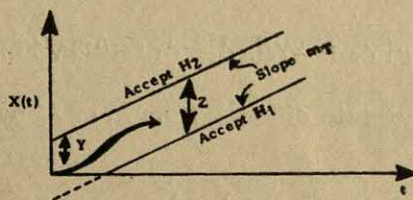


Figure 5.

Next, the Laplace transforms of the functions F and G of (2.13) and (2.16) are :

$$\int_0^{\infty} e^{-sy} F(y, x) dy = \frac{1}{s - x + x e^{-s}} \quad \dots (4.9)$$

and

$$\int_0^{\infty} e^{-sy} G(y, x) dy = \frac{x}{s(s - x + x e^{-s})}. \quad \dots (4.10)$$

MINIMAX TESTS FOR POISSON PROCESS AND NORMAL PROCESS BIAS RATES

Using (4.6) and letting $\gamma \rightarrow \infty$, the Laplace transforms of $\frac{1}{\sqrt{\gamma}}F$ and $\frac{1}{\gamma}G$ are easily inverted. From (2.12) and (2.15) we thus obtain :

$$P(2|1) = \frac{\exp[-2(m_T - m_1)Y/R] - \exp[-2(m_T - m_1)Z/R]}{1 - \exp[-2(m_T - m_1)Z/R]} \quad \dots \quad (4.11)$$

$$P(1|2) = \frac{\exp[2(m_2 - m_T)Y/R] - 1}{\exp[2(m_2 - m_T)Z/R] - 1} \quad \dots \quad (4.12)$$

$$\text{and} \quad i(i) = \frac{1}{m_T - m_i} \{ZP(1|i) - Y\}. \quad \dots \quad (4.13)$$

These formulas are not really new, being obtainable from Wald's treatment of the discrete problem of distinguishing between the means of two normal distributions. First, as in Dvoretzky, Kiefer and Wolfowitz (1953), with variance now R per unit time, the likelihood ratio is

$$L(t) = \exp \left\{ \frac{\Delta m}{R} [X(t) - mt] \right\}.$$

Comparison with (4.7) shows that the Wald parameters A and B are given by : $\ln A = Y\Delta m/R$ and $\ln B = -(Z - Y)\Delta m/R$. Next, (3.43) and (3.48) of Wald (1947), with θ_1 and θ_0 replaced by $m_2\Delta t$ and $m_1\Delta t$, and with θ replaced in turn by $m_1\Delta t$ and $m_2\Delta t$, yield (4.11) and (4.12). Finally, (4.13) is obtainable from (3.57) and (3.60) from Wald (1947), with σ^2 replaced by $R\Delta t$.

As pointed out in Dvoretzky, Kiefer and Wolfowitz (1953), Wald's formulas are exact in the case of the normal process, since there is no longer any overflow at the boundaries of the random walk.

Our only purpose in having derived these formulas here in a rather heuristic manner is to immediately extend our treatment to obtain the Bayes' and minimax procedures for distinguishing between the bias rates of two normal processes with the same diffusion rate R , when the losses W_{12} and W_{21} (due to a wrong decision) are known and when the cost of observation is c per unit time.

From the main result of Section 2, we now proceed to the main result of this Section :

For any positive values of W_{12} , W_{21} , c , R , and Δm , (a) there exist unique numbers ϕ , η_1 , η_2 , and Z , obtainable successively from :

$$\left\{ \frac{W_{21}(\Delta m)^2}{4cR} - \phi \right\} \left\{ \frac{W_{12}(\Delta m)^2}{4cR} - \phi \right\} = \sinh^2 \phi, \quad \dots (4.14)$$

$$\frac{\eta_1}{1-\eta_1} = \frac{\frac{W_{21}(\Delta m)^2}{4cR} + \frac{1}{2}(e^{2\phi} - 1) - \phi}{\frac{W_{12}(\Delta m)^2}{4cR} + \frac{1}{2}(1 - e^{-2\phi}) - \phi}, \quad \dots (4.15)$$

$$\frac{1-\eta_2}{\eta_2} = \frac{1-\eta_1}{\eta_1} e^{2\phi}, \quad \dots (4.16)$$

and
$$Z = \frac{2R\phi}{\Delta m}. \quad \dots (4.17)$$

(b) The Bayes' procedure Δ_η for distinguishing between the bias rates m_1 and m_2 , given an *a priori* weighting $\eta : 1-\eta$ in favour of m_1 , is as follows :

for $\eta \geq \eta_1$, accept H_1 outright; for $\eta \leq \eta_2$, accept H_2 outright;

for $\eta_2 < \eta < \eta_1$ apply the SPR test, defined by (4.7) and (4.8) or Figure 5, with Z determined from (a) and Y determined by :

$$Y = \frac{Z}{2\phi} \ln \frac{\eta(1-\eta_2)}{(1-\eta)\eta_2}. \quad \dots (4.18)$$

The minimized Bayes' risk $r(\eta, \Delta_\eta)$ again satisfies (3.1); and $r(\eta, \Delta)$, if restricted to SPR tests, is again stationary with respect to the SPR parameters, in this case m_T , Z , Y .

As in the case of the Poisson process, $Y \rightarrow Z$ as $\eta \rightarrow \eta_1$ from below and the test degenerates into outright acceptance of H_1 (see Figure 5 and/or formulas (4.11) through (4.13)).

Furthermore, $Y \rightarrow 0$ as $\eta \rightarrow \eta_2$ from above, and, in contrast with the Poisson problem, the test again degenerates, this time into outright acceptance of H_2 . This means that the curve of $r(\eta, \Delta_\eta)$ has a continuous slope at $\eta = \eta_2$ as well as elsewhere, so that Figure 3 is never applicable.

As in Figure 2, then, the minimax test procedure is an unmixed procedure, being the SPR test corresponding to the η^* and Y^* for which

$$r_1(m_T, Z, Y^*) = r_2(m_T, Z, Y^*). \quad \dots (4.19)$$

This, with the aid of (4.11), (4.12) and (4.13), provides an additional transcendental equation for the appropriate value Y^* of Y .

MINIMAX TESTS FOR POISSON PROCESS AND NORMAL PROCESS BIAS RATES

5. THE EXTENDED MINIMAX PROBLEM FOR A NORMAL PROCESS

Let us now turn to an important extension of the previous problem wherein the bias rate, say m , of a normal process can have any value and it is desired to decide whether m is above or below some critical rate m_0 . In other words, we are testing the hypothesis $H_1: m < m_0$ against the hypothesis $H_2: m > m_0$. The loss W_{21} due to accepting H_1 when H_2 is true will be, in general, a monotonically increasing function $W_{21}(m)$ of m defined over the range $m > m_0$; while the loss W_{12} due to accepting H_2 when H_1 is true will be, in general, a monotonically decreasing function $W_{12}(m)$ of m defined over the range $m < m_0$.

Let us first confine our attention to the three-parameter family of Wald-type tests defined by (4.7), the test parameters Y , Z , and m_T now being regarded as independent. The risk function r is given by the two formulas (2.3), according as $m \{ \leq \} m_0$, combined with the Wald formulas (4.11) through (4.13), with m_1 (now replaced by m) taking on all values less than m_0 , and m_2 (replaced by m) taking on all values greater than m_0 .

Assuming that the loss functions $W_{21}(m)$ and $W_{12}(m)$ vanish at the critical rate $m = m_0$ and increase continuously (and differentiably) but not too drastically as m moves away from m_0 (so that the expected losses $W_{21}(m)P(1|2)$ and $W_{12}(m)P(2|1)$ will actually decrease to zero again as $|m - m_0|$ increases indefinitely), the typical risk function has the general form of the full curve shown in Figure 6, in which the dotted curve represents the contribution $ci(m)$ of the cost of observation to the total risk r .

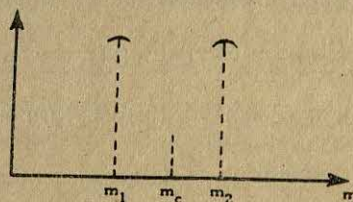


Figure 6.

The minimax choice among the three-parameter family of the Wald-type tests is that test which will render the maximum ordinate in Figure 6 as small as possible. Now we would naturally expect from Figure 5 that $P(1|2)$, for fixed Y and Z , would be an increasing function of m_T while $P(2|1)$ would be a decreasing function of m_T . These facts may be verified from (4.11) and (4.12). We may expect, moreover, that the maximum risk over the range $m > m_0$ would be a continuous increasing function of m_T , while the maximum risk over the range $m < m_0$ would be a continuous decreasing function of m_T . If so, the minimax choice of test parameters Y , Z , and m_T necessarily satisfies the condition:

$$r(m_1; m_T, Z, Y) = r(m_2; m_T, Z, Y), \text{ (as in Figure 6)} \quad \dots (5.1)$$

where

$$r(m_1; m_T, E, Y) = r_1(m_1; m_T, Z, Y) = \max_{m < m_0} r_1(m; m_T, Z, Y), \quad \dots (5.2)$$

$$r(m_2; m_T, E, Y) = r_2(m_2; m_T, Z, Y) = \max_{m > m_0} r_2(m; m_T, Z, Y),$$

Since r is, except at m_0 , a continuous and differentiable function of m , Y , Z and m_T , the minimization of r_1 and r_2 , subject to (5.1), requires the existence of a number η such that;

$$\eta r_1 + (1 - \eta) r_2 \quad \dots \quad (5.3)$$

is stationary with respect to the parameters m_T , Z , Y . But we have seen in Section 4 that this is the case for SPR tests belonging to the Bayes' procedures for distinguishing between the bias rates m_1 and m_2 (at which the risk function has its two equal maximums). Moreover, (5.1) is identical with (4.19), so that our necessary conditions (5.1) and (5.3) will both be satisfied by the minimax test for distinguishing between m_1 and m_2 .

The possibility that the minimax Wald-type test would not only lead to a risk function with two equal maximums, but would also be the minimax test for distinguishing between the bias rates at these maximum points was suggested to the author by both Herman Rubin and Melvin Peisakoff. This possibility, if realized, has the important consequence that the test in question is minimax among all possible test procedures, mixed or otherwise. For it follows from Section 4 that no other test procedure can have a lower over-all maximum.

Given c , m_0 and the loss functions $W_{21}(m)$ and $W_{12}(m)$, we are led to look for Wald-type test parameters m_T , Z and Y and for a real number ϕ which will satisfy not only (5.1) and (5.2) but also (4.8), (4.14) and (4.17). The test, if obtainable, will be minimax over the extended range of process rates among all possible test procedures.

This minimax solution has been carried out in the case of linear loss functions :

$$W_{12}(m) = A_1(m_0 - m), m < m_0; \quad W_{21}(m) = A_2(m - m_0), m > m_0; \quad \dots \quad (5.4)$$

where A_1 and A_2 are positive constants which are not necessarily equal. Unless they are equal, m_T is not equal to m_0 . Using (5.4), (4.14) provides a relation R_1 between the quantities ϕ , $\frac{A_1(\Delta m)^2(m_0 - m_1)}{cR}$, and $\frac{A_2(\Delta m)^2(m_2 - m_0)}{cR}$, while (5.1) using (2.3), (4.3), (4.8), (4.17) and (4.11) through (4.13), provides a relation R_2 between ϕ , Y/Z , $\frac{A_1(\Delta m)^2(m_0 - m_1)}{cR}$, $\frac{A_2(\Delta m)^2(m_2 - m_0)}{cR}$. Next, necessary conditions for the fulfilment of (5.2) are :

$$\frac{\partial r_1(m_1; m_T, Z, Y)}{\partial m_1} = 0 \quad \text{and} \quad \frac{\partial r_2(m_2; m_T, Z, Y)}{\partial m_2} = 0 \quad \dots \quad (5.5)$$

which, because of (2.3), (4.3), (4.8), (4.17) and (4.11) through (4.13) provide respectively a relation R_3 between ϕ , $\frac{Y}{Z}$, $\frac{A_1(\Delta m)^3}{cR}$ and $\frac{A_1(\Delta m)^2(m_0 - m_1)}{cR}$, and a relation R_4 between ϕ , $\frac{Y}{Z}$, $\frac{A_2(\Delta m)^3}{cR}$ and $\frac{A_2(\Delta m)^2(m_2 - m_0)}{cR}$. The four relations R_1 , R_2 , R_3 , and R_4 suffice

MINIMAX TESTS FOR POISSON PROCESS AND NORMAL PROCESS BIAS RATES

to determine $\frac{Y}{Z}$, $\frac{A_2}{A_1}$, $\frac{A_i}{cR}$ ($i = 1, 2$), and $\frac{m_0 - m_1}{\Delta m}$ as functions of the single parameter ϕ . Hence, using (4.8) and (4.17), the quantities

$$\left. \begin{aligned} \eta_R &= \left(\frac{c}{A_{av}R} \right)^{1/3} Y, \\ \eta_A &= \left(\frac{c}{A_{av}R} \right)^{1/3} (Z - Y), \\ \beta_I &= \left(\frac{A_{av}}{cR} \right)^{1/3} (m_T - m_0), \\ \theta_I &= \left(\frac{A_{av}}{cR} \right)^{1/3} (m_2 - m_T) = \left(\frac{A_{av}}{cR} \right)^{1/3} (m_T - m_1) \end{aligned} \right\} \quad \dots \quad (5.6)$$

as well as

$$\omega_I = \frac{r_{\max}}{(A_{av})^{2/3} (cR)^{1/3}} \quad \dots \quad (5.7)$$

in all of which

$$A_{av} = \frac{1}{2}(A_1 + A_2), \quad \dots \quad (5.8)$$

are obtainable as functions of the single economic parameter :

$$\rho = A_2/A_1. \quad \dots \quad (5.9)$$

Note that our formulation of the risk according to (2.3) and (5.4) involves four economic parameters : A_1 , A_2 , m_0 and c . The forms (5.6) and (5.7) of the minimax solution as a function of the single economic parameter ρ of (5.9) could, however, have been anticipated for "dimensional" reasons because of the invariance of the problem under the group of scale changes, not only of time and cost but also of the scale on which the normal process itself is measured (the vertical scale in Figure 5). The actual numerical determination of the quantities (5.6) and (5.7) as functions of ρ required the assistance of an IBM CPC. These are shown in Figure 7. As could be expected, the solution has a symmetry about $\rho = 1$, according to which the interchange of A_1 and A_2 is equivalent to the interchange of Y and $Z - Y$, together with a change in the sign of $m_T - m_0$.

This solution was given earlier by the author Breakwell (1954) in an application to the problem of testing for the fraction of defective, and was used again in a slightly different application on p. 255 of Breakwell (1956).

The minimax nature of the obtained solution will be established if we can prove that the necessary conditions (5.5) are also sufficient in combination with (4.8) and (4.14), to insure condition (5.2). The author has not succeeded in obtaining an analytical proof of this except in the economically symmetric case* $A_1 = A_2$, in which $m_T = m_0$, but numerical computations at a large number of values of ρ show that (5.2) is, indeed, satisfied.

* A separate account of this case has been given by Morris H. DeGroot.

6. THE EXTENDED MINIMAX PROBLEM FOR THE POISSON PROCESS

We turn now to the corresponding Poisson problem. In this case we are testing the hypothesis $H_1: \lambda < \text{some } \lambda_0$ against the hypothesis $H_2: \lambda > \lambda_0$, the loss functions W_{12} and W_{21} now being functions of λ , functions we again suppose to vanish at the critical rate (in this case λ_0) and to increase continuously (and differentiably) but not too drastically as λ moves away from λ_0 .

Guided by the results of Section 5, we look for a test procedure which not only leads to equal maximums on the risk curve at rates λ_1 and λ_2 , say, respectively below and above the critical rate λ_0 , but which is the minimax procedure for distinguishing between these two particular rates. In the light of Sections 2 and 3, this means one of two things: either

(1) The test procedure is an unmixed SPR test, with parameters μ and z related to λ_1 and λ_2 (according to (2.9) and (2.17)), and with parameter y related to λ_1 and λ_2 (by (3.3)); or

(2) It is a mixture in a ratio $\xi:1-\xi$ between outright acceptance of H_2 and an SPR test for which $y=0$ and for which μ and z are still related to λ_1 and λ_2 (according to (2.9) and (2.17)). Here ξ is related to λ_1 and λ_2 by (3.4).

For the mixed strategy, the risk is defined by:

$$r(\lambda) = \begin{cases} \xi W_{12}(\lambda) + (1-\xi)\{W_{12}(\lambda)P(2|\lambda; \mu, z, 0) + c\bar{t}(\lambda; \mu, z, 0)\}, & \text{if } \lambda < \lambda_0; \\ (1-\xi)\{W_{21}(\lambda)P(1|\lambda; \mu, z, 0) + c\bar{t}(\lambda; \mu, z, 0)\}, & \text{if } \lambda > \lambda_0; \end{cases} \quad \dots \quad (6.1)$$

whereas for an unmixed SPR test the risk is:

$$r(\lambda) = \begin{cases} W_{12}(\lambda)P(2|\lambda; \mu, z, y) + c\bar{t}(\lambda; \mu, z, y), & \text{if } \lambda < \lambda_0 \\ W_{21}(\lambda)P(1|\lambda; \mu, z, y) + c\bar{t}(\lambda; \mu, z, y), & \text{if } \lambda > \lambda_0 \end{cases} \quad \dots \quad (6.2)$$

In either case, the test procedure, if obtainable, will be minimax over the extended range of Poisson rates among all possible test procedures.

This minimax solution has been carried out in the case of loss functions of the same linear type as that treated in the extended normal problem, namely:

$$\begin{cases} W_{12}(\lambda) = K_1(\lambda_0 - \lambda), & \lambda < \lambda_0; \\ W_{21}(\lambda) = K_2(\lambda - \lambda_0), & \lambda > \lambda_0; \end{cases} \quad \dots \quad (6.3)$$

where K_1 and K_2 are positive constants which are not necessarily equal. (Even when K_1 and K_2 are equal, the minimax choice of test slope μ does not coincide with λ_0 because of the lack of symmetry of the Poisson distribution).

Relations are obtainable from Sections 2 and 3 and the locally maximum character of the risk function $r(\lambda)$ at λ_1 and λ_2 , which are sufficient to determine y or $\xi, z, \frac{\lambda_1}{\lambda_0}, \frac{\lambda_2}{\lambda_0}$, and hence $\frac{\mu}{\lambda_0}$ and $\frac{r_{\max}}{\sqrt{cK_2}}$ as functions of the two economic parameters:

$$\rho = K_2/K_1, \quad \text{and} \quad \delta = \lambda_0 \sqrt{\frac{K_2}{c}}. \quad \dots \quad (6.4)$$

MINIMAX TESTS FOR POISSON PROCESS AND NORMAL PROCESS BIAS RATES

This general form for the minimax solution could have been anticipated for dimensional reasons because of the invariance of the problem under the group of scale changes of time and cost. The scale of the Poisson process itself (the ordinate scale in Figure 1), is, of course, an intrinsically unique scale.

The numerical determination of this solution was carried out with the aid of an IBM 701 over the range $.01 < \rho < 100$. An unmixed minimax solution is obtained in this way at all "economic points" (ρ, δ) lying above a certain boundary on which the minimax test has $y = 0$. This boundary is the heavy line in Figure 8 and the upper dotted line in Figure 9. Below the line the minimax solution is a mixed strategy with $\xi > 0$ but $y = 0$. Figure 8 shows the values of ξ over the mixed region as well as the values of the ratio:

$$\alpha = \frac{y}{z} \quad \dots \quad (6.5)$$

over that part of the unmixed region for which $z \geq 1$.

As noted at the end of Section 2, those "limited" SPR tests for which $z < 1$ are more simply described by a single parameter T , the upper limit to observation time. The quantity:

$$v = T \sqrt{\frac{c}{K_2}} \quad \dots \quad (6.6)$$

is shown in Figure 9 over the relevant region. Above this region is shown the parameter z . The lower dotted line in Figures 9 and 10 is the upper boundary of a region for which the left maximum of $r(\lambda)$ occurs at $\lambda_1 = 0$ with $r'(0) \leq 0$.

Figure 10 shows the minimized maximum risk through the ratio:

$$\omega = \frac{r_{\max}}{\sqrt{cK_2}} \quad \dots \quad (6.7)$$

and the minimax test parameter μ , if $z \geq 1$, through the quantity:

$$\beta = \delta^{-1/3}(\lambda_0 - \mu) \sqrt{\frac{K_2}{c}} \quad \dots \quad (6.8)$$

As in the case of the extended normal problem, the minimax nature of the obtained solutions will be established if we can prove that the local maximum character of the risk function in the neighbourhood of λ_1 and λ_2 , in combination with (2.9) and (2.17), is sufficient to insure that the risk function does assume its highest value at λ_1 and λ_2 . Again, this has not been proven, but numerical computations at a large number of economic points (ρ, δ) indicate that this is so.

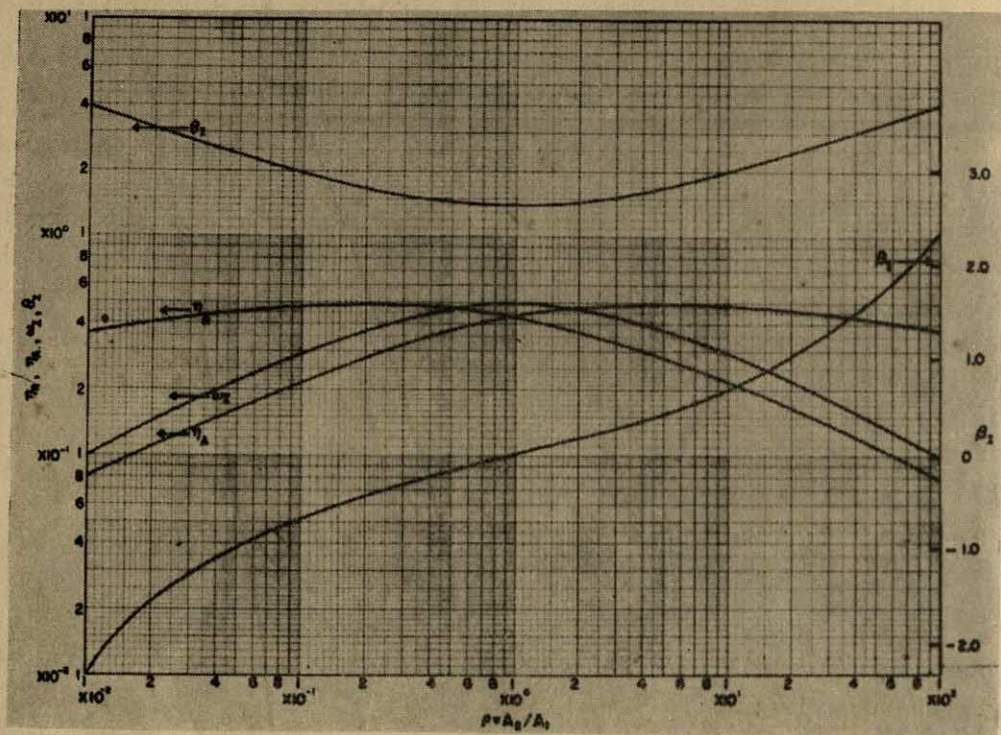


Figure 7.

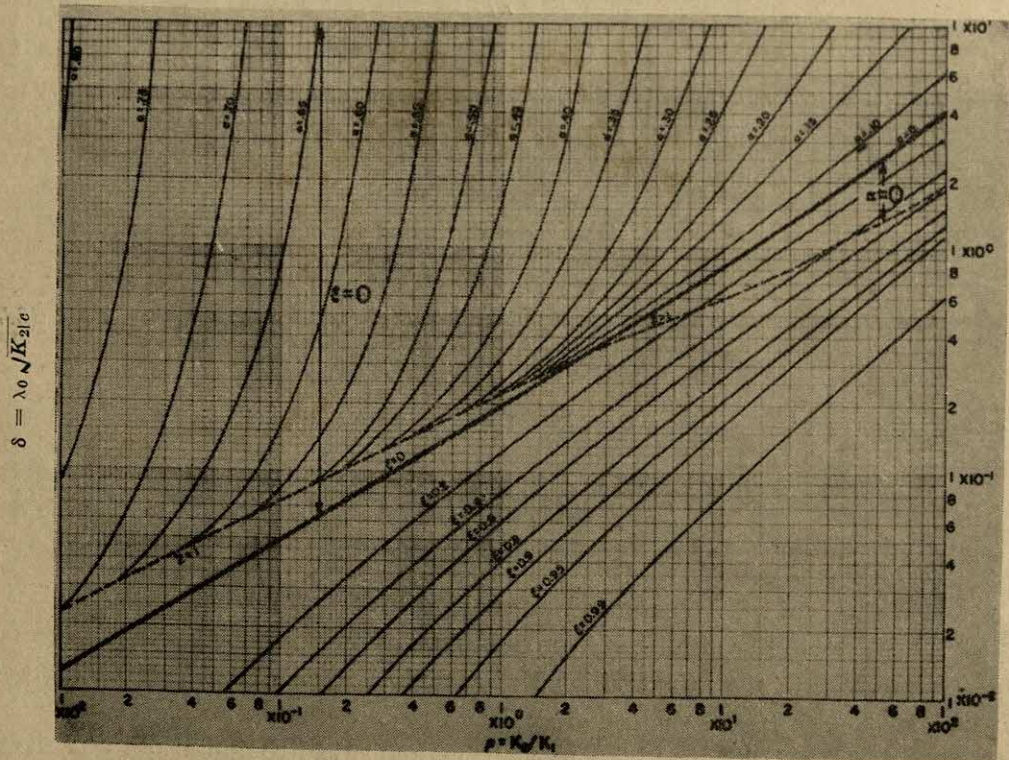


Figure 8.

MINIMAX TESTS FOR POISSON PROCESS AND NORMAL PROCESS BIAS RATES

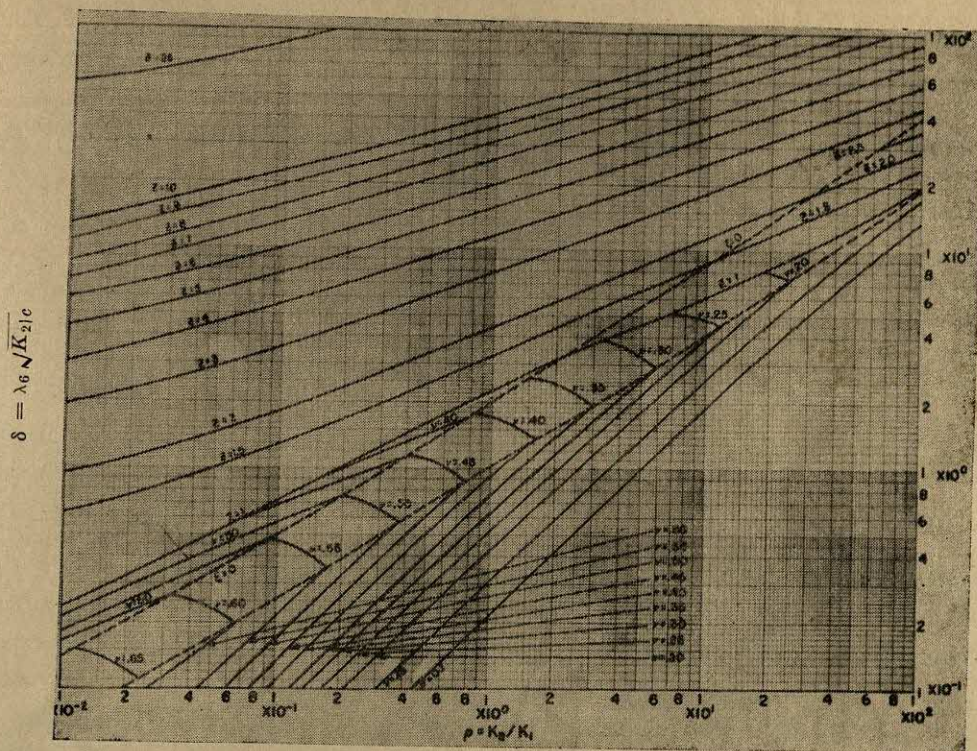


Figure 9.

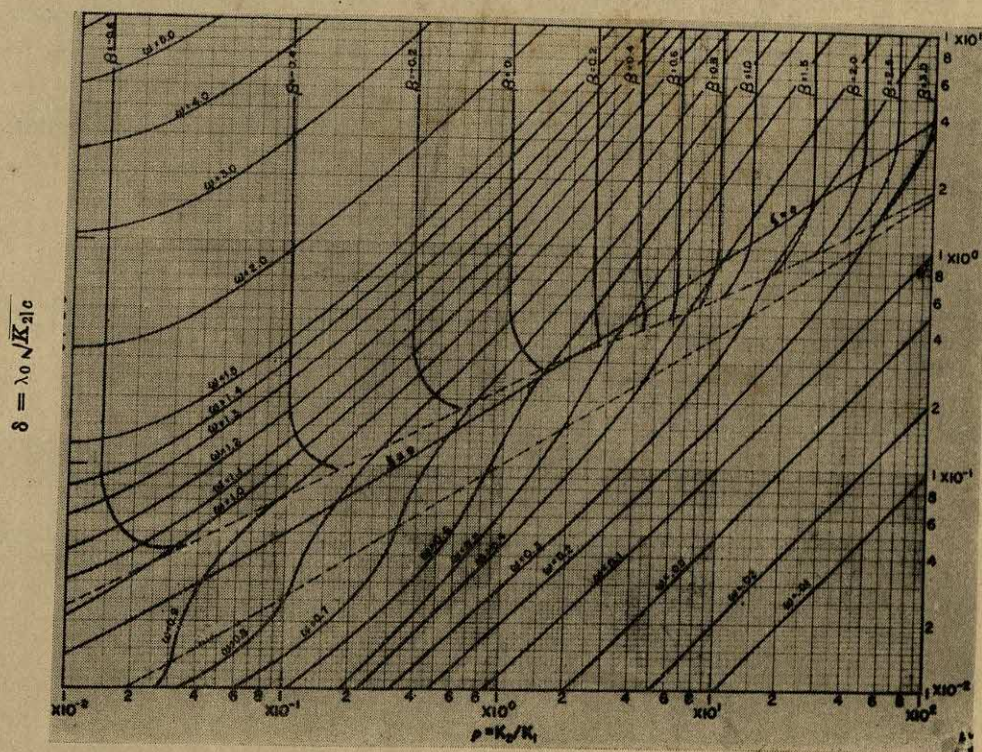


Figure 10.

REFERENCES

- BLACKWELL, D. and GIRSHICK, M. A. (1954): *Theory of Games and Statistical Decisions*, John Wiley and Sons, Inc., New York.
- BREAKWELL, J. V. (1954): The problem of testing for the fraction of defectives. *Journal of the Operations Research Society of America*, **2**(1), 59-69.
- (1956): Economically optimum acceptance tests. *J. Amer. Stat. Ass.*, **51**, 243-256.
- BURMAN, J. P. (1946): Sequential sampling for a Binomial population. *J. Roy. Stat. Soc.*, Suppl. **8**, 98-103.
- CHANDRASEKHAR, S. (1943): Stochastic problems in physics and astronomy. *Rev. Mod. Phys.*, **15**, 1.
- DVORETSKY, A., KIEFER, J. and WOLFOWITZ, J. (1953): Sequential decision problems for processes with continuous time parameter; testing hypotheses. *Ann. Math. Stat.*, **24**, 254-264.
- HAMMERSLEY, J. M. (1953): On counters with random dead time I. *Proc. Camb. Phil. Soc.*, **49**, 623-637.
- MORIGUTI, S. (1955): Notes on sampling inspection plans. *Rep. Stat. Appl. Res.*, JUSE, **3**(4), 99-121.
- UHLENBECK, G. E. and WANG, M. C. (1945): On the theory of Brownian motion II. *Rev. Mod. Phys.*, **17**, 323.
- WALD, A. (1950): *Statistical Decision Functions*, John Wiley and Sons, Inc., New York.
- (1947): *Sequential Analysis*, John Wiley and Sons, Inc., New York.

Paper received : March, 1959.

MINIMUM VARIANCE UNBIASED ESTIMATION IN COIN TOSSING PROBLEMS

By T. V. NARAYANA and Y. S. SATHE

University of Alberta and Summer Research Institute, Kingston

SUMMARY. The use of series expansions in estimation problems for distributions involving more than one parameter is illustrated in this paper. Unbiased minimum variance estimates are derived for the probabilities p_1, p_2 of obtaining heads with two coins 1, 2 in two "Markov" binomial sampling problems.

1. INTRODUCTION

We consider the problem of unbiased minimum variance (UMV) estimation in binomial sampling with two coins 1, 2 with probabilities p_1, p_2 of obtaining heads respectively ($p_1 \neq p_2$). Such problems have been considered by Narayana (1955), Narayana, Sathe and Chorneyko (1960) and the use of series expansions in UMV estimation problems has been discussed by DeGroot (1959), Guttman (1958). In Section 2, we consider an example of estimation in an inverse binomial sampling problem, which though artificial, illustrates these principles and generalises previous results. The generating function (g.f.) for this problem is given explicitly in Section 3, following methods given by Feller (1957). The rest of the paper deals with the application of similar combinatorial methods and the use of g.f.'s in related problems (Narayana and Ladouceur).

2. UMV ESTIMATION IN MARKOV INVERSE BINOMIAL SAMPLING

Consider the game N_r played with two coins 1, 2 with probabilities p_1, p_2 for heads ($0 < p_1 < 1; p_1 \neq p_2, p_1 \neq q_2 = 1 - p_2$) according to the following rules:

- (i) The first trial is made with coin 1.
- (ii) The n -th trial is made with coin 1 or 2 according as the $(n-1)$ -st trial is tail or head ($n > 1$).
- (iii) We stop the sequence of trials at that trial where the total number of heads equals r for the first time, r being a positive integer.

In what follows x and O stand for head and tail respectively.

The game N_r can end at the $(n+r)$ -th trial ($n = 0, 1, 2, \dots$) of a sequence of nO 's and rx 's, the last trial being, of course, x . Consider the sequences ending at the (n_0+r) -th trial which (a) either begin with an x and have $k-1$ changes of type Ox or (b) begin with a O and have k changes of type Ox . The probability associated

with such sequences is $p_1^k p_2^{r-k} q_1^{n-k+1} q_2^{k-1}$ and therefore (n, k) , by an appropriate choice of k for each n , can be considered a minimal sufficient statistic for the distribution. For given n , k varies between 1 and $\min(r, n+1)$. The number of sequences of type (a) i.e. starting with x and having $k-1$ changes, Ox are in one-to-one correspondence with the different possible arrangements of $r-2x$'s and nO 's with $k-1$ runs of O 's. The number of (a) sequences is thus

$$\binom{n-1}{k-2} \binom{r-1}{k-1}. \quad \dots (2.1)$$

Similarly the number of sequences (b) starting with exactly ν O 's ($\nu = 1, 2, \dots, n-k+1$) is

$$\binom{n-1-\nu}{k-2} \binom{r-1}{k-1} \quad \dots (2.2)$$

As the number of sequences in (2.1) is the same as substituting $\nu = 0$ in (2.2), the total number of sequences with probability $p_1^k p_2^{r-k} q_1^{n-k+1} q_2^{k-1}$, is

$$\begin{aligned} N'_r(k, n) &= \sum_{r=0}^{n-k+1} \binom{n-1-\nu}{k-2} \binom{n}{k-1} \\ &= \binom{r-1}{k-1} \binom{n}{k-1}. \end{aligned} \quad \dots (2.3)$$

Hence

$$\begin{aligned} &\sum_{n=0}^{\infty} \sum_{k=1}^{\min(r, n+1)} N'_r(k, n) p_1^k p_2^{r-k} q_1^{n-k+1} q_2^{k-1} \\ &= \sum_{n=0}^{\infty} \sum_{k=1}^{\min(r, n+1)} N'_r(k, n) \theta_1^k \theta_2^n m(\theta_1, \theta_2) = 1 \end{aligned} \quad \dots (2.4)$$

where $\theta_1 = \frac{p_1 q_2}{p_2 q_1}$, $\theta_2 = q_1$ and $m(\theta_1, \theta_2) = \frac{p_2^r q_1}{q_2} = \frac{(1-\theta_2 + \theta_1 \theta_2)^{r-1}}{\theta_1 (1-\theta_2)^{r-2}}$.

The form (2.4) indicates that the statistic (k, n) is complete. Therefore the unique UMV estimates of p_1, p_2 are

$$p_1 = \frac{\binom{n-1}{k-2}}{\binom{n}{k-1}} = \frac{k-1}{n}; \quad p_2 = \frac{\binom{r-2}{k-1}}{\binom{r-1}{k-1}} = \frac{r-1}{r-k}. \quad \dots (2.5)$$

MINIMUM VARIANCE UNBIASED ESTIMATION IN COIN TOSSING PROBLEMS

3. THE GENERATING FUNCTION OF $N'_r(s)$

Consider the game N_r which is played with the same rules as N'_r except that at the first trial coin 2 is used instead of coin 1 (cf (i), Section 2). Let $N'_r(s)$, $N_r(s)$ denote the g.f. of these games i.e.

$$N'_r(s) = \sum_{n=0}^{\infty} \sum_{k=1}^{\min(r, n+1)} N'_r(k, n) p_1^k p_2^{r-k} q_1^{n-k+1} q_2^k s^{n+2}. \quad \dots (3.1)$$

A method similar to Section 2 gives

$$N_r(s) = \sum_{n=0}^{\infty} \sum_{k=0}^{\min(r, n+1)} \binom{n}{k-1} \binom{r}{k} p_1^k p_2^{r-k} q_1^{n-k} q_2^k s^{n+r} \quad \dots (3.2)$$

where we make the convention that the coefficient $\binom{0}{-1} = 1$, $\binom{n}{-1} = 0$ for all positive n . Setting $N_0(s) = N'_0(s) = 1$, we have the relations

$$\left. \begin{aligned} N_r(s) &= p_2 s N_{r-1}(s) + q_2 s N'_r(s) \\ N'_r(s) &= p_1 s N_{r-1}(s) + q_1 s N'_r(s) \end{aligned} \right\} \quad \dots (3.3)$$

with
$$N'_1(s) = \frac{p_1 s}{1 - q_1 s} \text{ and } N_1(s) = p_2 s + \frac{p_1 q_2 s^2}{1 - q_1 s}. \quad \dots (3.4)$$

Using (3.3), (3.4) we see easily

$$N_r(s) = [N_1(s)]^r \text{ and } N'_r(s) = N'_1(s)[N_1(s)]^{r-1} \quad \dots (3.5)$$

a relation, which we expect to hold from the definitions of the games. Writing equations (3.1), (3.2) as power series in s , the relations (3.3) are valid. Inversely (3.4), (3.5) yield in turn (3.1), (3.2) and the UMV unbiased estimates for p_1 , p_2 . Differentiation of (3.5) using (3.4) gives the mean and variance of the number of trials required to terminate N_r , N'_r

$$E(N_r) = \frac{r(p_1 + q_2)}{p_1}, \quad E(N'_r) = (r-1) \frac{p_1 + q_2}{p_1} + \frac{1}{p_1}$$

4. GENERATING FUNCTIONS FOR RELATED PROBLEMS

Consider the games G_r , G'_r and their g.f.'s $G_r(s)$, $G'_r(s)$ defined analogously to N_r , N'_r except that the stopping rule specifies that the total number of heads exceeds the total number of tails by r for the first time [Narayana (1955); Narayana, Sathe and Chorneyko (1960)]. In order that the games G_r , G'_r terminate with probability 1 it is necessary and sufficient that $p_1 + p_2 \geq q_1 + q_2$. We assume this condition holds.

As

$$\begin{aligned} G_r(s) &= p_2 s G_{r-1}(s) + q_2 s G'_{r+1}(s) \\ G'_r(s) &= p_1 s G_{r-1}(s) + q_1 s G'_{r+1}(s). \end{aligned}$$

Where

$$G_0(s) = G'_0(s) = 1, \text{ we see as before} \\ G_r(s) = [G_1(s)]^r, G'_r(s) = G'_1(s)[G_1(s)]^{r-1} \quad \dots \quad (4.1)$$

$G_1(s)$, $G'_1(s)$ are in fact the negative roots of the equations :

$$\left. \begin{aligned} q_1 s x^2 - x\{1 + p_2 q_1 s^2 - p_1 q_2 s^2\} + p_2 s &= 0 \\ q_1 s x^2 - x\{1 + p_1 q_2 s^2 - p_2 q_1 s^2\} + p_1 s &= 0 \end{aligned} \right\} \quad \dots \quad (4.2)$$

Differentiating (4.2) using (4.1) we have means and variances of the number of trials required to terminate G_r , G'_r . These results generalise Narayana (1954). The expected values are

$$E(G_r) = r(p_1 + q_2)/(p_1 - q_2) \\ E(G'_r) = (r-1) \left(\frac{p_1 + q_2}{p_1 - q_2} \right) + \frac{p_2 + q_1}{p_1 - q_2}.$$

Their connection with Feller (1957) and the replacement of the two coins in G_r by a single coin with probabilities $\frac{p_1}{p_1 + q_2}$, $\frac{q_2}{p_1 + q_2}$ for heads and tails is clear. The problem of obtaining unbiased estimates for p_1 , p_2 in G_r , G'_r has been treated by Narayana, Sathe and Chorneyko (1960) and these results in turn imply the identity

$$\sqrt{1 + u^2 + v^2 - 2u - 2v - 2uv} = 1 - u - v - 2 \sum_{s=1}^{\infty} \sum_{t=1}^{\infty} \frac{\binom{s+t-1}{s} \binom{s+t-1}{t}}{(s+t-1)} u^s v^t \quad \dots \quad (4.3)$$

We are investigating an independent proof of (4.3) by analytic methods. We finally remark that (4.3) and the use of g.f.'s simplifies similar problems in unbiased estimation, (cf. Narayana and Ladouceur).

ACKNOWLEDGEMENT

We are most grateful to Professor R. L. Jeffery and the Summer Research Institute, Kingston, where this paper was written.

REFERENCES

- DEGROOT, M. H. (1959): Unbiased sequential estimation for binomial populations. *Ann. Math. Stat.*, **30**, 80.
- FELLER, W. (1957): *An Introduction to the Theory of Probability and Its Applications*, Vol. 1, John Wiley & Sons, Inc. (New York).
- GUTTMAN, I. (1958): A note on a series solution of a problem in estimation. *Biometrika*, **45**, 565.
- NARAYANA, T. V. (1954): A problem in the theory of probability. *J. Indian Soc. Agric. Stat.*, **6**, 139.
- (1955): Sur les treillis formés par les partitions d'un entier; leurs applications à la théorie des probabilités, CR t. 240, 1188.
- NARAYANA, T. V., MOHANTY, S. G. and LADOCEUR, J. D.: A combinatorial problem and its application to probability theory-II. (To appear in *J. Indian Soc. Agric. Stat.*).
- NARAYANA, T. V., SATHE, Y. S. and CHORNEYKO, I. (1960): Sufficient partitions for a class of coin-tossing problems. *Biometrische Zeitschrift*, Band 2, Heft 4.

Paper received : May, 1960.

CONFLICTING CRITERIA OF 'GOODNESS' OF STATISTICS

By J. SETHURAMAN

Indian Statistical Institute

SUMMARY. In this note it is intended to show by an example that if B and C are two statistics unbiased for a parameter θ and C has uniformly smaller variance than B , then for testing $\theta = \theta_0$ against an alternative (say $\theta > \theta_0$) the test based on C need not have more power than B locally (i.e., near the null hypothesis).

Basu (1953) gave an example (also quoted by Rao (1952)) to show that there can exist estimates with uniformly lesser variance than maximum likelihood estimates. We use the same example here.

$x_1, x_2, \dots, x_n$ are n independent observations from a rectangular distribution on $(\theta, 2\theta)$, $\theta > 0$. We wish to test the null hypothesis $\theta = \theta_0$ against the alternative $\theta > \theta_0$. Let us write

$$\xi = \max(x_1, x_2, \dots, x_n), \eta = \min(x_1, x_2, \dots, x_n), t = \frac{\eta}{\xi}.$$

Consider the statistics

$$A = (\xi, \eta) \text{ or equivalently } (\xi, t); \quad B = \xi; \quad C = 2\xi + \eta.$$

It is easy to verify that A is sufficient for θ but not complete; B is the maximum likelihood estimate for θ ; $(n+1)C/(5n+4)$ is the minimum variance unbiased linear estimator of ξ and η for θ , and

$$V\left(\frac{n+1}{5n+4} C\right) = \frac{1}{(n+2)(5n+4)} \theta^2$$

$$E\left(\frac{n+1}{2n+1} B\right) = \theta, \quad V\left(\frac{n+1}{2n+1} B\right) = \frac{n}{(n+2)(2n+1)^2} \theta^2$$

which show that, though B is the maximum likelihood estimate, there is another statistic with uniformly smaller variance.

Based on these statistics we can find (by an application of the Neyman-Pearson Lemma) uniformly most powerful tests A , B and C respectively for testing $\theta = \theta_0$, against $\theta > \theta_0$. The tests A , B and C of size α and their power functions β_A , β_B and β_C are given below as functions of $\lambda = \frac{\theta}{\theta_0}$.

Test A. Reject the null hypothesis if $\eta = \xi t \geq m$ and/or $\xi > 2\theta_0$

$$\mu = \frac{m}{\theta_0} = 2 - \alpha^{1/n}$$

$$\beta_A(\lambda) \left\{ \begin{array}{ll} = 1 - \frac{1}{\lambda^n} [(2-\lambda)^n - \alpha] & \text{if } 1 \leq \lambda \leq \mu \\ = 1 & \text{if } \mu < \lambda \end{array} \right\} \quad \dots \quad (1)$$

Test B. Reject the null hypothesis if $\xi \geq m$

$$\mu = \frac{m}{\theta_0} = 1 + (1 - \alpha)^{1/n}$$

$$\beta_B(\lambda) \left\{ \begin{array}{ll} = 1 - \frac{1}{\lambda^n} [1 + (1 - \alpha)^{1/n} - \lambda]^n & \text{if } 1 \leq \lambda \leq \mu \\ = 1 & \text{if } \mu < \lambda \end{array} \right\} \quad \dots \quad (2)$$

Test C. Reject the null hypothesis if $2\xi + \eta \geq m$

$$\mu = \frac{m}{\theta_0} = 6 - (3\alpha)^{1/n}$$

$$\beta_C(\lambda) \left\{ \begin{array}{ll} = \frac{[6\lambda - 6 + (3\alpha)^{1/n}]^n}{3\lambda^n} & \text{if } 1 \leq \lambda \leq \frac{\mu}{5} \\ = 1 - \frac{[6 - (3\alpha)^{1/n} - 3\lambda]^n}{3 \cdot 2^{n-1} \lambda^n} & \text{if } \frac{\mu}{5} \leq \lambda \leq \frac{\mu}{3} \\ = 1 & \text{if } \frac{\mu}{3} < \lambda \end{array} \right\} \quad \dots \quad (3)$$

The numerical values of the powers of these tests (for the size 0.05) are given in the following Table for some selected values of λ and n .

The test A based on the sufficient statistic is naturally better than the other two tests. But it is surprising to find that test B is more powerful than test C in the neighbourhood of θ_0 . The following Table shows that test C is poorer than test B in the neighbourhood of the null hypothesis, even though test C catches up and exceeds test B in power for higher values of λ . The practical statistician may note that in this region of large values of λ the power of both the tests is quite high that one would not be much the worse by using the poorer test B.

CONFLICTING CRITERIA OF GOODNESS OF STATISTICS

TABLE. POWER FUNCTION

null hypothesis: $\theta = \theta_0$ i.e., $\lambda = 1$; alternative: $\theta > \theta_0$ i.e., $\lambda > 1$; size of the test = 0.05

λ/n	test based on (ξ, η) the sufficient statistic					test based on ξ the m.l.e.					test based on $2\xi + \eta$ the u.m.v.u. linear estimate of ξ and η for θ .				
	1	2	5	10	20	1	2	5	10	20	1	2	5	10	20
1.0	.0500	.0500	.0500	.0500	.0500	.0500	.0500	.0500	.0500	.0500	.0500	.0500	.0500	.0500	.0500
1.1	.2273	.3719	.6644	.8848	.9894	.2273	.3677	.6537	.8732	.9828	.2273	.2993	.5695	.8668	.9892
1.2	.3750	.5903	.8884	.9907	1.0000	.3750	.5832	.8765	.9837	.9997	.3750	.5322	.8755	.9900	1.0000*
1.3	.5000	.7396	.9682	1.0000	1.0000	.5000	.7306	.9579	.9981	1.0000*	.5000	.7207	.9680	.9995	1.0000*
1.4	.6071	.8418	.9948	1.0000	1.0000	.6071	.8315	.9867	.9998	1.0000*	.6071	.8303	.9933	1.0000*	1.0000*
1.5	.7000	.9111	1.0000	1.0000	1.0000	.7000	.8999	.9963	1.0000*	1.0000*	.7000	.9082	.9990	1.0000*	1.0000*
1.6	.7812	.9570	1.0000	1.0000	1.0000	.7812	.9452	.9991	1.0000*	1.0000*	.7812	.9570	1.0000*	1.0000*	1.0000*
1.7	.8529	.9862	1.0000	1.0000	1.0000	.8529	.9739	.9999	1.0000*	1.0000*	.8529	.9849	1.0000*	1.0000*	1.0000
1.8	.9167	1.0000	1.0000	1.0000	1.0000	.9167	.9894	1.0000*	1.0000*	1.0000*	.9167	.9979	1.0000	1.0000	1.0000
1.9	.9737	1.0000	1.0000	1.0000	1.0000	.9737	.9985	1.0000*	1.0000*	1.0000*	.9737	1.0000	1.0000	1.0000	1.0000
2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

* correct to four decimal places.

Differentiating the power function at $\lambda = 1$ we have

$$\left. \begin{aligned} \beta'_A(1) &= n(2-\alpha) \\ \beta'_B(1) &= n((1-\alpha)^{\frac{n-1}{n}} + 1 - \alpha) \sim n(2-2\alpha) \\ \beta'_C(1) &= n(2.(3\alpha)^{\frac{n-1}{n}} - \alpha) \sim n.5\alpha \end{aligned} \right\} \dots (4)$$

This clearly shows that test B is locally more powerful than test C.

ACKNOWLEDGEMENT

I wish to thank Professor C. R. Rao for suggesting this investigation to me.

REFERENCES

- BASU, D. (1953): Unpublished D.Phil. Thesis: Some contributions to the theory of inference: submitted to Calcutta University.
- RAO, C. R. (1952): Minimum variance estimates in distributions admitting ancillary statistics, *Sankhyā*, **12**, 53-56.

Paper received : April, 1960.

MISCELLANEOUS

SAMPLING EXPERIMENTS ON THE COMBINATION OF INDEPENDENT χ^2 -TESTS

By NIKHILESH BHATTACHARYA

Indian Statistical Institute

SUMMARY. The present communication reports on some model sampling work on the relative powers of three methods of combining independent χ^2 -tests. One is based on the straightforward addition of χ^2 's and of the corresponding degrees of freedom; the second is the P_{λ_n} technique using upper tail probabilities associated with the χ^2 's; and the third is Tippett's test (Birnbaum, 1954) based on the minimum upper tail probability. Although the experiments are in no way exhaustive, they indicate that the first two methods are almost equally powerful, (which has interesting implications,) and that the third method is usually inferior to the other two.

1. OUTLINE OF THE EXPERIMENT

1.1. Let k denote the d.f. and λ the parameter of non-centrality for a non-central χ^2 . For each of a number of combinations of k and λ , one series of independent non-central χ^2 's was built up by using Wold's table of normal deviates (1948). The three methods of combination were then applied* to mutually exclusive sets of n non-central χ^2 's of each series, n being, in turn, 2, 6, 12 or 24. Table 1 summarises the results.

1.2. Another type of experiments was carried out for χ^2 's with single d.f. Two series of single d.f. χ^2 's were taken, the two differing in respect of λ , and a 'mixed' series was built up by picking up alternate elements from the two 'pure' series. The three methods of combination were then applied to mutually exclusive sets of n non-central χ^2 's of the 'mixed' series, where $n = 2, 6, 12$ or 24 . Results for such 'mixed' series are shown in Table 2.

1.3. Power figures given in the tables are all 'estimates' based on the model sampling work, although 'true' values were also calculated for the first and the third methods using Patnaik's approximate rules (1949). These 'true' values agreed with the corresponding estimates to within limits of sampling error. The 'estimates' are, however, presented here instead of 'true' values, in the interest of making the power comparisons more sensitive.

1.4. To save space, estimates of power are given for two particular levels of significance. Results for the other levels were very similar. Also, lines for $n = 12$ or 24 are omitted in case the number of experiments fall below 50.

2. RESULTS

2.1. As regards the relative powers of $\Sigma\chi^2$ and P_{λ_n} tests, Table 1 shows that these are almost equally efficient when the χ^2 's combined have the same k and λ , where $k = 1, 6, 12$ or 24 , and λ assumes the common range of values. Table 2 indicates that some variation in λ does not alter the situation if $k = 1$. For $k = 2$, it may be recalled, the two methods are strictly equivalent, whether the λ 's are different or not. From all these, it becomes probable that the two methods are almost equally powerful even in the general case of combining χ^2 's with varying k and λ .

*Upper tail probabilities (q) were, of course, easily found for χ^2 's with single degrees of freedom. For higher d.f., formula (22) given by Pearson and Hartley (1958, Introduction, pp.13-14) was used when $q > 0.001$; when $q < 0.001$, Tables of the Incomplete Gamma Function (K. Pearson, 1946) were used.

TABLE 1. RELATIVE POWERS OF THREE METHODS OF COMBINING n INDEPENDENT χ^2 -TESTS WHEN k AND λ ARE EQUAL FOR ALL THE χ^2 's

parameters of the individual non-central χ^2 's combined		number of χ^2 's combined (n)	number of model experi- ments	estimated powers (%)					
				at 5% level			at 0.1% level		
				$\Sigma\chi^2$ test	P_{λ_n} test	Tippett's test	$\Sigma\chi^2$ test	P_{λ_n} test	Tippett's test
k	λ			(5)	(6)	(7)	(8)	(9)	(10)
1	0.25	2	600	7.3	7.7	6.7	0.7	0.7	0.3
		6	200	12.0	11.5	9.0	—	—	—
		12	100	14.0	17.0	11.0	3.0	3.0	—
		24	50	18.0	20.0	10.0	4.0	4.0	—
1	1.0	2	600	21.2	21.7	20.0	1.3	1.7	1.0
		6	200	39.0	37.5	24.0	7.0	7.5	1.5
		12	100	51.0	54.0	23.0	17.0	17.0	1.0
		24	50	82.0	84.0	34.0	28.0	30.0	—
1	2.25	2	600	46.2	46.7	40.5	6.8	7.7	3.8
		6	200	82.0	82.0	54.5	31.5	34.0	6.5
		12	100	97.0	100.0	64.0	66.0	69.0	6.0
		24	50	100.0	100.0	68.0	100.0	100.0	8.0
1	4.0	2	600	72.3	73.2	64.8	22.2	22.8	12.3
		6	200	98.0	98.5	84.0	78.0	78.5	15.5
		12	100	100.0	100.0	95.0	100.0	100.0	23.0
		24	50	100.0	100.0	100.0	100.0	100.0	32.0
1	6.25	2	600	90.5	91.2	85.2	48.3	45.3	28.2
		6	200	100.0	100.0	97.0	97.0	98.0	46.5
		12	100	100.0	100.0	100.0	100.0	100.0	55.0
		24	50	100.0	100.0	100.0	100.0	100.0	70.0
6	1.5	2	600	17.3	17.3	13.5	1.7	1.5	0.8
		6	200	24.5	23.5	17.5	3.5	4.0	1.0
		12	100	39.0	38.0	17.0	8.0	8.0	2.0
		24	50	68.0	68.0	22.0	22.0	22.0	2.0
6	6.0	2	600	59.7	58.3	51.5	17.2	15.8	7.7
		6	200	95.0	91.5	73.0	60.0	58.5	11.5
		12	100	100.0	100.0	82.0	95.0	93.0	15.0
		24	50	100.0	100.0	90.0	100.0	100.0	24.0
12	3.0	2	300	21.7	22.0	18.3	3.0	2.7	3.3
		6	100	39.0	36.0	29.0	8.0	9.0	1.0
		12	50	68.0	68.0	32.0	22.0	22.0	—
12	12.0	2	300	82.7	80.3	77.7	38.3	37.3	23.0
		6	100	100.0	100.0	91.0	95.0	93.0	37.0
		12	50	100.0	100.0	98.0	100.0	100.0	52.0
24	6.0	2	150	34.0	33.3	29.3	4.7	4.7	4.7
		6	50	68.0	70.0	44.0	22.0	24.0	8.0
24	24.0	2	150	98.3	98.0	96.0	76.0	76.0	57.3
		6	50	100.0	100.0	100.0	100.0	100.0	82.0

SAMPLING EXPERIMENTS ON THE COMBINATION OF INDEPENDENT χ^2 -TESTS

TABLE 2. RELATIVE POWERS OF THREE METHODS OF COMBINING n INDEPENDENT SINGLE D.F. χ^2 -TESTS, WHEN THE λ 's ARE NOT EQUAL FOR ALL THE χ^2 's

values of λ for the χ^2 's combined		number of χ^2 's combined (n)	number of model experi- ments	estimated powers (%)					
				at 5% level			at 0.1% level		
				$\Sigma\chi^2$ test	P_{λ_n} test	Tippett's test	$\Sigma\chi^2$ test	P_{λ_n} test	Tippett's test
for one set of $\frac{n}{2}$ χ^2 's	for the other set of $\frac{n}{2}$ χ^2 's			(5)	(6)	(7)	(8)	(9)	(10)
0	1	2	600	13.5	13.8	14.2	1.0	0.8	0.5
		6	200	22.0	20.5	17.0	3.0	2.0	1.0
		12	100	25.0	26.0	17.0	4.0	4.0	1.0
		24	50	36.0	38.0	24.0	10.0	10.0	2.0
0.25	4	2	600	43.7	42.8	42.2	6.8	6.0	7.3
		6	200	79.0	81.0	61.5	28.5	27.0	8.5
		12	100	99.0	99.0	77.0	60.0	61.0	15.0
		24	50	100.0	100.0	86.0	98.0	98.0	22.0
0	6.25	2	600	61.8	61.2	63.2	15.0	14.2	16.0
		6	200	95.5	95.0	84.0	54.0	51.0	28.0
		12	100	100.0	100.0	96.0	98.0	97.0	36.0
		24	50	100.0	100.0	100.0	100.0	100.0	42.0
1	2.25	2	600	33.0	33.2	30.5	3.8	3.8	3.0
		6	200	61.5	62.5	38.5	16.0	17.5	5.0
		12	100	93.0	93.0	45.0	34.0	36.0	4.0
		24	50	100.0	100.0	50.0	74.0	84.0	4.0
1	4	2	600	52.3	52.8	48.3	9.5	9.7	6.3
		6	200	87.0	86.5	65.0	35.0	37.5	8.5
		12	100	98.0	97.0	78.0	80.0	81.0	9.0
		24	50	100.0	100.0	88.0	98.0	98.0	14.0
2.25	6.25	2	600	74.7	74.8	69.5	24.2	24.5	18.2
		6	200	97.5	97.5	90.0	82.0	82.5	28.0
		12	100	100.0	100.0	96.0	98.0	99.0	35.0
		24	50	100.0	100.0	98.0	100.0	100.0	50.0

2.2. Tippett's method seems to be less efficient than the other two in most cases, although for $n = 2$, the difference is small or sometimes even zero. The difference increases with n and is more marked for the 0.1% level. It is, however, possible that when one or a few of the λ 's is sufficiently higher than the rest, the method may even be superior to the other two.

2.3. That the addition method is more efficient than the (min q) method is indirectly seen from Table 1 : the (min q) method applied to a set of non-central χ^2 's becomes more efficient when applied to sub-totals of the same χ^2 's.

3. FURTHER OBSERVATIONS

3.1. Combination of χ^2 's may become necessary when one has carried out a series of goodness of fit tests or tests on a number of contingency tables, and where just one test on the pooled data may not be meaningful or adequate. In some cases there may be need of giving unequal weights to the tests being combined, (Yates, 1955a, 1955b; Zelen, 1957), but this point has been ignored in the present study.

3.2. The result for single d.f. χ^2 's seems to have interesting implications, and in what follows, only single d.f. χ^2 's are considered. Let $x_1, x_2, \dots, x_n$ be independent normally distributed variates, each having unit s.d., but with $E(x_i) = \mu_i$; and let us suppose that the μ_i 's are free to have any sign and magnitude. Let y_i be the incomplete probability integral corresponding to x_i , calculated with reference to the standard normal distribution. Then to test the hypothesis $H(\mu_1 = \mu_2 = \dots = \mu_n = 0)$, one can use either $\sum_{i=1}^n x_i^2$ or the Sukhatme form of $P_{\lambda_n} = \prod_{i=1}^n z_i$, where $z_i = 1 - 2|y_i - \frac{1}{2}|$. Since z_i is the upper tail probability corresponding to x_i^2 , which is a non-central χ^2 with 1 d.f., Tables 1 and 2 imply that these two tests are of nearly equal power, although $\sum x_i^2$ is generally believed to have some optimum properties for this well-known model which has direct bearing on the combination of independent two-sided tests, and hence to tests of homogeneity.

3.3. The two criteria are, however, closely similar. Whereas $\sum x_i^2$ is a sum of single d.f. χ^2 's, $-2 \log_e (P_{\lambda_n})$ is the sum $\sum (-2 \log_e z_i)$, and $-2 \log_e z_i$ is that value of χ^2 with 2 d.f. which corresponds to x_i^2 in having the same incomplete probability integral y_i .

3.4. Birnbaum (1954) considered this model for the simple case $n = 2$, and found that the critical regions defined by the two criteria are very similar.* Earlier, Lancaster (1949) had studied the problem of combining two-sided tests on 2×2 tables or on binomial data, for the case where small frequencies are involved; and his work seemed to suggest that, for combining the single d.f. χ^2 's, the summation method and the P_{λ_n} technique (using upper tail areas) would be about equally powerful.

3.5. Earlier still, E. S. Pearson (1938) had showed that the critical region given by low values of $P_{\lambda_n} = \prod_{i=1}^n [1 - 2|y_i - \frac{1}{2}|]$ is optimum for testing whether a sample of x -values ($x_1, x_2, \dots, x_n$) has probably arisen from a population $N(0, 1)$, where the alternative hypothesis states that x is $N(0, \sigma)$, with $\sigma > 1$, y_i being the incomplete probability integral of x_i under

*As regards Tippet's test Birnbaum's (1954) recommendations were largely influenced by his consideration for the heterogeneous case. For (more or less) homogeneous cases even criteria leading to non-convex acceptance regions may be definitely superior to Tippet's test. Even for ordinary heterogeneous cases, Tippet's test will become comparatively inefficient as n increases beyond 2. This is obviously because, unlike the P_{λ_n} -test or the $\sum \chi^2$ -test, Tippet's test is unduly dependent on one extreme observation.

SAMPLING EXPERIMENTS ON THE COMBINATION OF INDEPENDENT χ^2 -TESTS

the null hypothesis. This result was not entirely correct, but it suggested that the Sukhatme form of the P_{λ_n} test is almost as efficient as the UMP test based on $\sum x_i^2$. It is interesting to note that this model is a first approximation to that mentioned in para 3.2.

3.6. Yates (1955b) remarked that the problems considered by Lancaster (1949) were unrealistic. It seems, however, that the problem of combining two-sided tests may arise with binomial data.

3.7. Suppose one is given some (large-sample) binomial data, arranged group-wise, and wants to test for a preassigned proportion p_0 for all the groups, where the group proportions can individually exceed or fall short of p_0 . Mather's monograph (1951, pp. 15-20) shows the application of χ^2 -tests to such problems. The P_{λ_n} test could equally be used in such cases, and might even be adapted to give approximate analysis of the total divergence into components for 'deviation' and 'heterogeneity'.

4. HOMOGENEITY OF CORRELATION COEFFICIENTS

4.1. One may next consider a situation where one has k sample correlation coefficients $r_1, r_2, \dots, r_k$, based on independent random samples from k bivariate normal populations. Let the respective sample sizes be $n_1, n_2, \dots, n_k$, and the population correlation coefficients be $\rho_1, \rho_2, \dots, \rho_k$; and suppose it is desired to test the hypothesis $H(\rho_1 = \rho_2 = \dots = \rho_k = \rho_0)$, ρ_0 being a preassigned value, where the ρ_i 's may either exceed or fall below ρ_0 individually.

4.2. The Fisherian test based on the z -transformation is well-known. K. Pearson (1933) suggested an alternative method based on probability integrals, but this was not properly oriented, and David (1938, pp. xxii-xxviii) rightly modified the Pearson test.

Let $p_i = \int_{-1}^{r_i} P(r/\rho_0, n_i) dr$, where $P(r/\rho_0, n_i)$ is the frequency function of r_i under the null hypothesis. Then the Pearson-David criterion is $\prod_{i=1}^n [1 - 2|p_i - \frac{1}{2}|]$, small values of the product being significant.

4.3. Consider the case where the n_i 's are so large that $y_i = \sqrt{n_i - 3}(z_i - \xi_0)$ can be regarded as standard normal deviates under the null hypothesis, where $z_i = \tanh^{-1} r_i$, and $\xi_0 = \tanh^{-1} \rho_0$. If now one notes that p_i is the probability integral of y_i also, the problem is seen to be equivalent to that considered in para 3.2. Fisher's criterion $\sum y_i^2$ corresponding to that based on $\sum x_i^2$, and the Pearson-David criterion to P_{λ_n} of that para. Theoretical considerations suggest that the Fisherian test would have some optimum properties; but it involves approximations, while the other test is exact, and as far as the present investigation can show, the differences in power are almost negligible in most cases.

4.4. There could be many other instances where the Sukhatme form of P_{λ_n} can be applied to test whether a number of unknown parameters $\theta_1, \theta_2, \dots, \theta_k$ are simultaneously equal to a preassigned value θ_0 . For a strict test of homogeneity, however, θ_0 should be left unspecified. In such cases, the parameter θ_0 has to be estimated from sample data before carrying out homogeneity tests and the exact distribution of P_{λ_n} becomes unknown. It is customary to still regard $-2 \log_e (P_{\lambda_n})$ as a χ^2 with $2k$ d.f., but this number $2k$ is obviously too high.

ACKNOWLEDGEMENTS

The author is glad to acknowledge the help received from Shri Rabindranath Mukherjee and Shri Bimalendu Mahalanobis who organised the model sampling experiments.

REFERENCES

- BERNBAUM, ALLAN (1954) : Combining independent tests of significance. *J. Amer. Stat. Ass.*, **49**, 559-574.
- DAVID, F. N. (1938) : *Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples*, First Edition, Cambridge University Press.
- LANCASTER, H. O. (1949) : The combination of probabilities arising from data in discrete distributions. *Biometrika*, **36**, 370-382.
- MATHER, K. (1951) : *The Measurement of Linkage in Heredity*, Second Edition, Methuen & Co. Ltd., London.
- PATNAIK, P. B. (1949) : The non-central χ^2 and F -distributions and their applications. *Biometrika*, **36**, 202-232.
- PEARSON, E. S. (1938) : The probability integral transformation for testing goodness of fit and combining independent tests of significance. *Biometrika*, **30**, 134-148.
- and HARTLEY, H. O. (1958) : *Biometrika Tables for Statisticians*, Vol. 1, Second Edition, Cambridge University Press.
- PEARSON, KARL (1933) : On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, **25**, 379-410.
- (1946) : *Tables of the Incomplete Γ -function*, Second Re-issue, Cambridge University Press.
- WOLD, H. O. A. (1948) : *Random Normal Deviates : Tracts for Computers*, No. XXV, Cambridge University Press.
- YATES, FRANK (1955a) : The use of transformations and maximum likelihood in the analysis of quantal experiments involving two treatments. *Biometrika*, **42**, 382-403.
- (1955b) : A note on the application of the combination of probabilities test to a set of 2×2 tables. *Biometrika*, **42**, 404-411.
- ZELEN, MARVIN (1957) : The analysis of incomplete block designs. *J. Amer. Stat. Ass.*, **52**, 204-217.

Paper received : February, 1960.

A TABLE OF THE NORMAL INTEGRAL WITH COMPLEX ARGUMENT

By PETER IHM

Botanic Institute, University of Freiburg, Germany

The function $\text{Erf}(z) = (2/\pi)^{1/2} \int_0^z e^{-\xi^2} d\xi$ is known under the name Error Function.

It has been studied extensively by Rosser (1948). For certain statistical purposes we require

$$\phi(z) = \lim_{a \rightarrow \infty} \frac{1}{\sqrt{2\pi}} \int_{-a}^z e^{-\frac{1}{2}\xi^2} d\xi, \quad -\frac{\pi}{4} \leq \arg a \leq \frac{\pi}{4}, \quad \dots (1)$$

which may be called Normal Integral. We have for $z = x + iy$

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}y^2} \left\{ \int_{-\infty}^x \cos \xi y e^{-\frac{1}{2}\xi^2} d\xi - i \int_{-\infty}^x \sin \xi y e^{-\frac{1}{2}\xi^2} d\xi \right\}.$$

The integrand in (1) may be expanded in a power series and yields after integration, term by term, a confluent hypergeometric series.

$$\phi^*(z) = \lim_{b \rightarrow \infty} \frac{1}{\sqrt{2\pi}} \int_z^b e^{-\frac{1}{2}\xi^2} d\xi, \quad -\frac{\pi}{4} \leq \arg b \leq \frac{\pi}{4},$$

gives for $z \neq 0$ the asymptotic expansion

$$\phi^*(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \left(\frac{1}{z} - \frac{1}{z^3} + \frac{1.3}{z^5} - \frac{1.3.5}{z^7} + \dots \right)$$

with absolute error smaller than the amount of the first neglected term. $\phi(z)$ has been calculated for $x = 0(0.1)\dots$, $y = 0(0.1)2.5$ by aid of an IBM Magnetic Drum Calculator Type 650* using

$$\text{Re } \phi(z) = \frac{1}{2} + \frac{e^{\frac{1}{2}y^2}}{\sqrt{2\pi}} \int_0^x \cos \xi y e^{\frac{1}{2}\xi^2} d\xi, \quad \text{Im } \phi(z) = \frac{1}{\sqrt{2\pi}} \left\{ \int_0^y e^{\frac{1}{2}\xi^2} d\xi - e^{\frac{1}{2}y^2} \int_0^{|x|} \sin \xi y e^{-\frac{1}{2}\xi^2} d\xi \right\}.$$

The following relations hold:

$$\text{Re } \phi(-z) = 1 - \text{Re } \phi(z), \quad \text{Im } \phi(-z) = -\text{Im } \phi(z), \quad \text{Re } \phi(\bar{z}) = \text{Re } \phi(z), \quad \text{Im } \phi(\bar{z}) = -\text{Im } \phi(z),$$

$$\phi(-z) = 1 - \phi(z), \quad \phi(\bar{z}) = \bar{\phi}(z);$$

$$\lim_{z \rightarrow \infty} \phi(-z) = 0, \quad \lim_{z \rightarrow \infty} \phi(z) = 1 \quad \text{for } -\frac{1}{4}\pi \leq \arg z \leq \frac{1}{4}\pi,$$

and

$$\phi(iy) = \frac{1}{2} + \frac{i}{\sqrt{2\pi}} \int_0^y e^{\frac{1}{2}\xi^2} d\xi.$$

* The work was supported by Deutsche Forschungsgemeinschaft. I am indebted to Professor A. Walther, Technische Hochschule Darmstadt, for the kind permission to use the electronic computer.

SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS : SERIES A

TABLES OF $\operatorname{Re} \phi(z)$ AND $\operatorname{Im} \phi(z)$; $z = x + iy$

x	$y = 0.0$		$y = 0.1$		$y = 0.2$		$y = 0.3$	
	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$
0.0	0.5000	0.0000	0.5000	0.0400	0.5000	0.0803	0.5000	0.1215
0.1	0.5398	0.0000	0.5400	0.0398	0.5406	0.0799	0.5417	0.1209
0.2	0.5793	0.0000	0.5797	0.0392	0.5808	0.0787	0.5829	0.1190
0.3	0.6179	0.0000	0.6185	0.0382	0.6202	0.0767	0.6232	0.1160
0.4	0.6554	0.0000	0.6562	0.0369	0.6584	0.0741	0.6622	0.1119
0.5	0.6915	0.0000	0.6923	0.0353	0.6950	0.0708	0.6995	0.1068
0.6	0.7257	0.0000	0.7267	0.0334	0.7298	0.0669	0.7349	0.1009
0.7	0.7580	0.0000	0.7591	0.0313	0.7624	0.0627	0.7681	0.0944
0.8	0.7881	0.0000	0.7893	0.0290	0.7928	0.0581	0.7988	0.0874
0.9	0.8159	0.0000	0.8171	0.0266	0.8208	0.0533	0.8269	0.0800
1.0	0.8413	0.0000	0.8426	0.0242	0.8462	0.0484	0.8524	0.0726
1.1	0.8643	0.0000	0.8655	0.0218	0.8692	0.0435	0.8753	0.0651
1.2	0.8849	0.0000	0.8861	0.0194	0.8896	0.0387	0.8955	0.0579
1.3	0.9032	0.0000	0.9043	0.0171	0.9077	0.0341	0.9133	0.0509
1.4	0.9192	0.0000	0.9203	0.0149	0.9235	0.0298	0.9287	0.0443
1.5	0.9332	0.0000	0.9342	0.0129	0.9371	0.0257	0.9420	0.0381
1.6	0.9452	0.0000	0.9461	0.0111	0.9488	0.0220	0.9532	0.0325
1.7	0.9554	0.0000	0.9562	0.0094	0.9586	0.0186	0.9626	0.0274
1.8	0.9641	0.0000	0.9648	0.0079	0.9669	0.0156	0.9705	0.0229
1.9	0.9713	0.0000	0.9719	0.0065	0.9738	0.0129	0.9769	0.0189
2.0	0.9772	0.0000	0.9778	0.0054	0.9794	0.0106	0.9821	0.0155
2.1	0.9821	0.0000	0.9826	0.0044	0.9840	0.0086	0.9862	0.0125
2.2	0.9861	0.0000	0.9865	0.0035	0.9876	0.0069	0.9896	0.0100
2.3	0.9893	0.0000	0.9896	0.0028	0.9906	0.0055	0.9922	0.0080
2.4	0.9918	0.0000	0.9921	0.0022	0.9929	0.0043	0.9942	0.0062
2.5	0.9938	0.0000	0.9940	0.0017	0.9947	0.0034	0.9957	0.0048
2.6	0.9953	0.0000	0.9955	0.0013	0.9960	0.0026	0.9969	0.0037
2.7	0.9965	0.0000	0.9967	0.0010	0.9971	0.0020	0.9978	0.0028
2.8	0.9974	0.0000	0.9976	0.0008	0.9979	0.0015	0.9984	0.0021
2.9	0.9981	0.0000	0.9982	0.0006	0.9985	0.0011	0.9989	0.0016
3.0	0.9987	0.0000	0.9987	0.0004	0.9989	0.0008	0.9992	0.0012
3.1	0.9990	0.0000	0.9991	0.0003	0.9992	0.0006	0.9995	0.0009
3.2	0.9993	0.0000	0.9994	0.0002	0.9995	0.0004	0.9996	0.0006
3.3	0.9995	0.0000	0.9995	0.0002	0.9996	0.0003	0.9998	0.0004
3.4	0.9997	0.0000	0.9997	0.0001	0.9997	0.0002	0.9998	0.0003
3.5	0.9998	0.0000	0.9998	0.0001	0.9998	0.0002	0.9999	0.0002
3.6	0.9998	0.0000	0.9999	0.0001	0.9999	0.0001	0.9999	0.0002
3.7	0.9999	0.0000	0.9999	0.0000	0.9999	0.0001	1.0000	0.0001
3.8	0.9999	0.0000	0.9999	0.0000	0.9999	0.0001	1.0000	0.0001
3.9	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000

A TABLE OF THE NORMAL INTEGRAL WITH COMPLEX ARGUMENT

TABLES OF $\operatorname{Re} \phi(z)$ AND $\operatorname{Im} \phi(z)$; $z = x + iy$

x	$y = 0.4$		$y = 0.5$		$y = 0.6$		$y = 0.7$	
	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$
0.0	0.5000	0.1639	0.5000	0.2081	0.5000	0.2545	0.5000	0.3038
0.1	0.5431	0.1631	0.5451	0.2070	0.5477	0.2531	0.5508	0.3021
0.2	0.5858	0.1605	0.5897	0.2036	0.5947	0.2489	0.6009	0.2968
0.3	0.6274	0.1563	0.6331	0.1982	0.6404	0.2420	0.6496	0.2882
0.4	0.6677	0.1507	0.6750	0.1908	0.6843	0.2326	0.6960	0.2766
0.5	0.7061	0.1437	0.7148	0.1817	0.7259	0.2211	0.7398	0.2623
0.6	0.7423	0.1356	0.7522	0.1711	0.7647	0.2078	0.7804	0.2459
0.7	0.7761	0.1266	0.7868	0.1595	0.8005	0.1931	0.8175	0.2278
0.8	0.8073	0.1170	0.8186	0.1470	0.8330	0.1775	0.8508	0.2085
0.9	0.8357	0.1069	0.8473	0.1340	0.8620	0.1612	0.8802	0.1886
1.0	0.8612	0.0967	0.8729	0.1209	0.8876	0.1448	0.9057	0.1686
1.1	0.8840	0.0866	0.8954	0.1078	0.9099	0.1286	0.9275	0.1490
1.2	0.9040	0.0767	0.9150	0.0951	0.9289	0.1130	0.9458	0.1300
1.3	0.9213	0.0672	0.9318	0.0830	0.9449	0.0981	0.9607	0.1121
1.4	0.9362	0.0583	0.9460	0.0717	0.9581	0.0842	0.9727	0.0955
1.5	0.9489	0.0500	0.9578	0.0612	0.9689	0.0714	0.9822	0.0804
1.6	0.9595	0.0425	0.9676	0.0517	0.9775	0.0599	0.9893	0.0668
1.7	0.9682	0.0357	0.9754	0.0432	0.9842	0.0497	0.9946	0.0549
1.8	0.9754	0.0297	0.9817	0.0357	0.9894	0.0407	0.9984	0.0445
1.9	0.9812	0.0244	0.9866	0.0291	0.9932	0.0330	1.0009	0.0356
2.0	0.9858	0.0198	0.9904	0.0236	0.9960	0.0264	1.0025	0.0282
2.1	0.9894	0.0160	0.9933	0.0188	0.9980	0.0209	1.0033	0.0220
2.2	0.9922	0.0127	0.9955	0.0149	0.9993	0.0163	1.0037	0.0169
2.3	0.9943	0.0100	0.9970	0.0116	1.0002	0.0126	1.0036	0.0129
2.4	0.9959	0.0078	0.9981	0.0090	1.0006	0.0096	1.0034	0.0097
2.5	0.9971	0.0060	0.9989	0.0069	1.0009	0.0073	1.0030	0.0071
2.6	0.9980	0.0046	0.9994	0.0052	1.0010	0.0054	1.0026	0.0052
2.7	0.9987	0.0035	0.9997	0.0039	1.0009	0.0040	1.0022	0.0037
2.8	0.9991	0.0026	0.9999	0.0029	1.0009	0.0029	1.0018	0.0027
2.9	0.9994	0.0019	1.0000	0.0021	1.0007	0.0021	1.0014	0.0018
3.0	0.9996	0.0014	1.0001	0.0015	1.0006	0.0015	1.0011	0.0013
3.1	0.9998	0.0010	1.0001	0.0011	1.0005	0.0010	1.0009	0.0009
3.2	0.9999	0.0007	1.0001	0.0008	1.0004	0.0007	1.0007	0.0006
3.3	0.9999	0.0005	1.0001	0.0005	1.0003	0.0005	1.0005	0.0004
3.4	1.0000	0.0004	1.0001	0.0004	1.0002	0.0003	1.0004	0.0002
3.5	1.0000	0.0003	1.0001	0.0002	1.0002	0.0002	1.0003	0.0001
3.6	1.0000	0.0002	1.0001	0.0002	1.0001	0.0001	1.0002	0.0001
3.7	1.0000	0.0001	1.0000	0.0001	1.0001	0.0001	1.0001	0.0001
3.8	1.0000	0.0001	1.0000	0.0001	1.0001	0.0001	1.0001	0.0000
3.9	1.0000	0.0001	1.0000	0.0000	1.0000	0.0000	1.0001	0.0000
4.0	1.0000	0.0000	—	—	—	—	1.0000	0.0000

SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS : SERIES A

TABLES OF $\operatorname{Re} \phi(z)$ AND $\operatorname{Im} \phi(z)$; $z = x + iy$

x	$y = 0.8$		$y = 0.9$		$y = 1.0$		$y = 1.1$	
	$\operatorname{Re} \phi(x)$	$\operatorname{Im} \phi(x)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$
0.0	0.5000	0.3567	0.5000	0.4140	0.5000	0.4767	0.5000	0.5460
0.1	0.5548	0.3545	0.5596	0.4113	0.5656	0.4734	0.5728	0.5420
0.2	0.6087	0.3480	0.6182	0.4034	0.6298	0.4637	0.6440	0.5302
0.3	0.6608	0.3375	0.6747	0.3905	0.6915	0.4480	0.7121	0.5110
0.4	0.7105	0.3232	0.7281	0.3731	0.7496	0.4268	0.7757	0.4852
0.5	0.7569	0.3058	0.7778	0.3518	0.8031	0.4010	0.8338	0.4539
0.6	0.7996	0.2857	0.8230	0.3274	0.8513	0.3715	0.8854	0.4182
0.7	0.8382	0.2636	0.8634	0.3007	0.8937	0.3393	0.9301	0.3795
0.8	0.8724	0.2402	0.8986	0.2725	0.9300	0.3055	0.9675	0.3391
0.9	0.9022	0.2162	0.9287	0.2437	0.9603	0.2711	0.9977	0.2981
1.0	0.9276	0.1921	0.9537	0.2150	0.9846	0.2371	1.0211	0.2579
1.1	0.9487	0.1685	0.9739	0.1871	1.0035	0.2042	1.0381	0.2195
1.2	0.9659	0.1460	0.9897	0.1605	1.0174	0.1733	1.0495	0.1835
1.3	0.9795	0.1248	1.0015	0.1359	1.0269	0.1447	1.0561	0.1507
1.4	0.9900	0.1054	1.0099	0.1133	1.0328	0.1189	1.0586	0.1215
1.5	0.9976	0.0878	1.0154	0.0932	1.0355	0.0961	1.0580	0.0960
1.6	1.0030	0.0721	1.0186	0.0754	1.0360	0.0763	1.0551	0.0742
1.7	1.0065	0.0585	1.0199	0.0602	1.0347	0.0595	1.0506	0.0560
1.8	1.0086	0.0468	1.0199	0.0472	1.0322	0.0455	1.0452	1.0412
1.9	1.0095	0.0369	1.0189	0.0364	1.0289	0.0340	1.0393	0.0293
2.0	1.0096	0.0287	1.0173	0.0277	1.0253	0.0249	1.0333	0.0202
2.1	1.0091	0.0220	1.0153	0.0206	1.0216	0.0178	1.0277	0.0132
2.2	1.0083	0.0166	1.0132	1.0151	1.0180	0.0123	1.0225	0.0081
2.3	1.0073	0.0123	1.0111	0.0108	1.0147	0.0082	1.0180	0.0045
2.4	1.0063	0.0090	1.0091	0.0075	1.0118	0.0052	1.0141	0.0021
2.5	1.0053	0.0064	1.0074	0.0051	1.0093	0.0031	1.0108	0.0005
2.6	1.0043	0.0045	1.0059	0.0034	1.0072	0.0017	1.0081	-0.0004
2.7	1.0034	0.0031	1.0046	0.0021	1.0054	0.0008	1.0060	-0.0009
2.8	1.0027	0.0021	1.0035	0.0013	1.0041	0.0002	1.0043	-0.0011
2.9	1.0021	0.0014	1.0026	0.0007	1.0030	-0.0001	1.0031	-0.0012
3.0	1.0016	0.0009	1.0019	0.0004	1.0021	-0.0003	1.0021	-0.0011
3.1	1.0012	0.0006	1.0014	0.0002	1.0015	-0.0004	1.0014	-0.0009
3.2	1.0009	0.0003	1.0010	0.0000	1.0010	-0.0003	1.0010	-0.0008
3.3	1.0006	0.0002	1.0007	0.0000	1.0007	-0.0003	1.0006	-0.0006
3.4	1.0004	0.0001	1.0005	-0.0001	1.0005	-0.0003	1.0004	-0.0005
3.5	1.0003	0.0000	1.0003	-0.0001	1.0003	-0.0002	1.0002	-0.0003
3.6	1.0002	0.0000	1.0002	-0.0001	1.0002	-0.0002	1.0001	-0.0002
3.7	1.0001	0.0000	1.0001	-0.0001	1.0001	-0.0001	1.0001	-0.0002
3.8	1.0001	0.0000	1.0001	0.0000	1.0001	-0.0001	1.0000	-0.0001
3.9	1.0001	0.0000	1.0001	0.0000	1.0000	-0.0001	1.0000	-0.0001
4.0	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000	-0.0001
4.1	—	—	—	—	—	—	1.0000	0.0000

A TABLE OF THE NORMAL INTEGRAL WITH COMPLEX ARGUMENT

TABLES OF $\operatorname{Re} \phi(z)$ AND $\operatorname{Im} \phi(z)$; $z = x + iy$

x	$y = 1.2$		$y = 1.3$		$y = 1.4$		$y = 1.5$	
	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$
0.0	0.5000	0.6234	0.5000	0.7106	0.5000	0.8100	0.5000	0.9242
0.1	0.5816	0.6185	0.5925	0.7046	0.6058	0.8025	0.6222	0.9151
0.2	0.6613	0.6040	0.6825	0.6868	0.7084	0.7807	0.7405	0.8880
0.3	0.7371	0.5806	0.7677	0.6581	0.8051	0.7454	0.8512	0.8445
0.4	0.8074	0.5492	0.8461	0.6198	0.8933	0.6985	0.9511	0.7867
0.5	0.8709	0.5112	0.9160	0.5736	0.9708	0.6420	1.0377	0.7174
0.6	0.9266	0.4680	0.9762	0.5213	1.0364	0.5785	1.1094	0.6400
0.7	0.9737	0.4214	1.0261	0.4652	1.0892	0.5106	1.1652	0.5578
0.8	1.0122	0.3730	1.0655	0.4072	1.1291	0.4411	1.2052	0.4742
0.9	1.0421	0.3244	1.0946	0.3495	1.1567	0.3725	1.2303	0.3924
1.0	1.0639	0.2771	1.1141	0.2937	1.1729	0.3068	1.2417	0.3151
1.1	1.0784	0.2321	1.1251	0.2413	1.1791	0.2459	1.2415	0.2443
1.2	1.0865	0.1906	1.1288	0.1936	1.1771	0.1911	1.2319	0.1817
1.3	1.0892	0.1532	1.1266	0.1511	1.1685	0.1433	1.2151	0.1280
1.4	1.0876	0.1203	1.1198	0.1144	1.1551	0.1027	1.1935	0.0836
1.5	1.0828	0.0921	1.1098	0.0836	1.1387	0.0694	1.1692	0.0483
1.6	1.0758	0.0684	1.0978	0.0583	1.1207	0.0429	1.1439	0.0213
1.7	1.0675	0.0491	1.0849	0.0383	1.1024	0.0228	1.1192	0.0017
1.8	1.0586	0.0338	1.0720	0.0230	1.0847	0.0081	1.0962	-0.0114
1.9	1.0496	0.0220	1.0595	0.0117	1.0684	-0.0020	1.0755	-0.0194
2.0	1.0411	0.0132	1.0482	0.0037	1.0539	-0.0084	1.0575	-0.0234
2.1	1.0334	0.0068	1.0381	-0.0015	1.0413	-0.0119	1.0425	-0.0245
2.2	1.0265	0.0025	1.0294	-0.0047	1.0309	-0.0133	1.0304	-0.0235
2.3	1.0206	-0.0003	1.0222	-0.0063	1.0224	-0.0133	1.0209	-0.0212
2.4	1.0157	-0.0020	1.0163	-0.0068	1.0158	-0.0123	1.0136	-0.0184
2.5	1.0117	-0.0028	1.0117	-0.0066	1.0107	-0.0108	1.0084	-0.0153
2.6	1.0085	-0.0030	1.0082	-0.0060	1.0069	-0.0091	1.0047	-0.0123
2.7	1.0060	-0.0030	1.0055	-0.0052	1.0043	-0.0074	1.0022	-0.0096
2.8	1.0042	-0.0027	1.0036	-0.0043	1.0024	-0.0059	1.0006	-0.0072
2.9	1.0028	-0.0023	1.0022	-0.0034	1.0012	-0.0045	0.9997	-0.0053
3.0	1.0019	-0.0019	1.0013	-0.0027	1.0004	-0.0033	0.9993	-0.0038
3.1	1.0012	-0.0015	1.0007	-0.0020	1.0000	-0.0024	0.9991	-0.0026
3.2	1.0007	-0.0011	1.0003	-0.0015	0.9998	-0.0017	0.9991	-0.0018
3.3	1.0004	-0.0009	1.0001	-0.0011	0.9997	-0.0012	0.9992	-0.0011
3.4	1.0002	-0.0006	1.0000	-0.0007	0.9997	-0.0008	0.9993	-0.0007
3.5	1.0001	-0.0004	0.9999	-0.0005	0.9997	-0.0005	0.9995	-0.0004
3.6	1.0000	-0.0003	0.9999	-0.0003	0.9998	-0.0003	0.9996	-0.0002
3.7	1.0000	-0.0002	0.9999	-0.0002	0.9998	-0.0002	0.9997	-0.0001
3.8	1.0000	-0.0001	0.9999	-0.0001	0.9999	-0.0001	0.9998	-0.0001
3.9	1.0000	-0.0001	0.9999	-0.0001	0.9999	-0.0001	0.9999	0.0000
4.0	1.0000	-0.0001	1.0000	-0.0001	0.9999	0.0000	0.9999	0.0000
4.1	1.0000	0.0000	1.0000	0.0000	0.9999	0.0000	0.9999	0.0000
4.2	—	—	—	—	1.0000	0.0000	1.0000	0.0000

SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS : SERIES A

TABLES OF $\operatorname{Re} \phi(z)$ AND $\operatorname{Im} \phi(z)$; $z = x + iy$

x	$y = 1.6$		$y = 1.7$		$y = 1.8$		$y = 1.9$	
	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$
0.0	0.5000	1.0571	0.5000	1.2129	0.5000	1.3977	0.5000	1.6190
0.1	0.6426	1.0456	0.6681	1.1986	0.7002	1.3797	0.7407	1.5960
0.2	0.7803	1.0120	0.8298	1.1565	0.8920	1.3266	0.9704	1.5288
0.3	0.9082	0.9579	0.9790	1.0890	1.0676	1.2419	1.1792	1.4216
0.4	1.0224	0.8864	1.1107	1.0000	1.2206	1.1304	1.3584	1.2812
0.5	1.1198	0.8111	1.2209	0.8943	1.3461	0.9988	1.5021	1.1162
0.6	1.1983	0.7062	1.3072	0.7775	1.4412	0.8542	1.6067	0.9362
0.7	1.2572	0.6062	1.3688	0.6554	1.5049	0.7042	1.6715	0.7512
0.8	1.2965	0.5054	1.4061	0.5334	1.5383	0.5560	1.6983	0.5704
0.9	1.3175	0.4078	1.4210	0.4166	1.5442	0.4159	1.6911	0.4018
1.0	1.3222	0.3166	1.4164	0.3090	1.5266	0.2889	1.6555	0.2517
1.1	1.3133	0.2345	1.3958	0.2138	1.4904	0.1787	1.5983	0.1245
1.2	1.2938	0.1632	1.3633	0.1329	1.4408	0.0874	1.5264	0.0221
1.3	1.2665	0.1034	1.3227	0.0670	1.3830	0.0155	1.4467	-0.0550
1.4	1.2346	0.0555	1.2777	0.0160	1.3218	-0.0375	1.3651	-0.1082
1.5	1.2005	0.0187	1.2317	-0.0211	1.2612	-0.0733	1.2868	-0.1403
1.6	1.1665	-0.0079	1.1873	-0.0459	1.2044	-0.0943	1.2153	-0.1547
1.7	1.1343	-0.0256	1.1464	-0.0603	1.1535	-0.1032	1.1532	-0.1553
1.8	1.1051	-0.0361	1.1103	-0.0665	1.1099	-0.1030	1.1016	-0.1459
1.9	1.0797	-0.0408	1.0798	-0.0664	1.0741	-0.0963	1.0608	-0.1300
2.0	1.0582	-0.0414	1.0549	-0.0621	1.0460	-0.0855	1.0301	-0.1108
2.1	1.0408	-0.0390	1.0353	-0.0552	1.0249	-0.0727	1.0084	-0.0906
2.2	1.0272	-0.0348	1.0207	-0.0470	1.0100	-0.0594	0.9942	-0.0712
2.3	1.0170	-0.0298	1.0102	-0.0386	1.0001	-0.0468	0.9860	-0.0537
2.4	1.0096	-0.0246	1.0032	-0.0305	0.9941	-0.0356	0.9822	-0.0388
2.5	1.0045	-0.0196	0.9988	-0.0233	0.9911	-0.0260	0.9814	-0.0268
2.6	1.0011	-0.0151	0.9963	-0.0172	0.9901	-0.0182	0.9826	-0.0175
2.7	0.9992	-0.0113	0.9953	-0.0123	0.9904	-0.0122	0.9849	-0.0106
2.8	0.9982	-0.0082	0.9951	-0.0084	0.9915	-0.0077	0.9876	-0.0058
2.9	0.9978	-0.0057	0.9955	-0.0055	0.9929	-0.0046	0.9903	-0.0025
3.0	0.9978	-0.0039	0.9961	-0.0035	0.9944	-0.0024	0.9927	-0.0006
3.1	0.9980	-0.0025	0.9968	-0.0020	0.9957	-0.0010	0.9948	0.0005
3.2	0.9983	-0.0016	0.9975	-0.0011	0.9969	-0.0002	0.9964	0.0010
3.3	0.9987	-0.0009	0.9982	-0.0005	0.9978	0.0002	0.9976	0.0012
3.4	0.9990	-0.0005	0.9987	-0.0001	0.9985	0.0004	0.9985	0.0011
3.5	0.9993	-0.0002	0.9991	0.0001	0.9990	0.0005	0.9991	0.0009
3.6	0.9995	-0.0001	0.9994	0.0001	0.9994	0.0004	0.9995	0.0007
3.7	0.9996	0.0000	0.9996	0.0002	0.9996	0.0004	0.9998	0.0006
3.8	0.9998	0.0000	0.9998	0.0002	0.9998	0.0003	0.9999	0.0004
3.9	0.9998	0.0001	0.9999	0.0001	0.9999	0.0002	1.0000	0.0003
4.0	0.9999	0.0000	0.9999	0.0001	1.0000	0.0001	1.0000	0.0002
4.1	0.9999	0.0000	1.0000	0.0001	1.0000	0.0001	1.0000	0.0001
4.2	1.0000	0.0000	1.0000	0.0000	1.0000	0.0001	1.0000	0.0001
4.3	—	—	—	—	1.0000	0.0000	1.0000	0.0000

A TABLE OF THE NORMAL INTEGRAL WITH COMPLEX ARGUMENT

TABLES OF $\operatorname{Re} \phi(z)$ AND $\operatorname{Im} \phi(z)$; $z = x + iy$

x	$y = 2.0$		$y = 2.1$		$y = 2.2$		$y = 2.3$	
	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$
0.0	0.5000	1.8866	0.5000	2.2135	0.5000	2.6168	0.5000	3.1195
0.1	0.7923	1.8573	0.8586	2.1757	0.9443	2.5678	1.0560	3.0554
0.2	1.0702	1.7714	1.1981	2.0652	1.3630	2.4245	1.5775	2.8681
0.3	1.3205	1.6348	1.5010	1.8899	1.7329	2.1979	2.0335	2.5729
0.4	1.5323	1.4567	1.7530	1.6623	2.0354	1.9051	2.3991	2.1932
0.5	1.6977	1.2485	1.9444	1.3980	2.2576	1.5671	2.6580	1.7580
0.6	1.8126	1.0232	2.0700	1.1142	2.3938	1.2072	2.8035	1.2987
0.7	1.8766	0.7938	1.8854	0.5834	2.4451	0.8485	2.8384	0.8463
0.8	1.8926	0.5723	2.1579	0.5842	2.4188	0.5116	2.7741	0.4281
0.9	1.8664	0.3689	2.4039	0.6373	2.3273	0.2133	2.6282	0.0658
1.0	1.8061	0.1915	2.6137	0.7316	2.1863	-0.0345	2.4230	-0.2264
1.1	1.7208	0.0449	2.7816	0.8546	2.0128	-0.2255	2.1820	-0.4415
1.2	1.6197	-0.0686	2.9057	0.9938	1.8238	-0.3586	1.9278	-0.5799
1.3	1.5119	-0.1494	1.5760	-0.2743	1.6345	-0.4373	1.6803	-0.6481
1.4	1.4051	-0.2003	1.4377	-0.3185	1.4571	-0.4685	1.4551	-0.6570
1.5	1.3052	-0.2251	1.3120	-0.3310	1.3007	-0.4613	1.2626	-0.6198
1.6	1.2167	-0.2290	1.2039	-0.3189	1.1707	-0.4259	1.1085	-0.5510
1.7	1.1421	-0.2173	1.1160	-0.2898	1.0691	-0.3723	0.9941	-0.4639
1.8	1.0824	-0.1952	1.0486	-0.2502	0.9953	-0.3095	0.9168	-0.3704
1.9	1.0372	-0.1671	1.0003	-0.2062	0.9466	-0.2449	0.8717	-0.2796
2.0	1.0052	-0.1369	0.9689	-0.1622	0.9189	-0.1839	0.8525	-0.1980
2.1	0.9843	-0.1075	0.9513	-0.1217	0.9077	-0.1302	0.8524	-0.1293
2.2	0.9725	-0.0808	0.9441	-0.0865	0.9083	-0.0858	0.8650	-0.0751
2.3	0.9675	-0.0579	0.9443	-0.0578	0.9166	-0.0511	0.8848	-0.0350
2.4	0.9672	-0.0393	0.9493	-0.0354	0.9290	-0.0256	0.9073	-0.0076
2.5	0.9699	-0.0248	0.9568	-0.0191	0.9429	-0.0081	0.9295	0.0094
2.6	0.9741	-0.0142	0.9651	-0.0078	0.9564	0.0028	0.9493	0.0183
2.7	0.9790	-0.0069	0.9732	-0.0006	0.9684	0.0087	0.9658	0.0215
2.8	0.9837	-0.0021	0.9804	0.0034	0.9784	0.0111	0.9786	0.0211
2.9	0.9880	0.0007	0.9864	0.0053	0.9861	0.0113	0.9880	0.0185
3.0	0.9915	0.0021	0.9910	0.0057	0.9918	0.0101	0.9943	0.0151
3.1	0.9943	0.0027	0.9945	0.0053	0.9957	0.0084	0.9983	0.0115
3.2	0.9963	0.0026	0.9969	0.0045	0.9982	0.0065	1.0006	0.0083
3.3	0.9978	0.0023	0.9985	0.0036	0.9997	0.0047	1.0017	0.0056
3.4	0.9988	0.0019	0.9994	0.0027	1.0005	0.0033	1.0020	0.0036
3.5	0.9994	0.0015	1.0000	0.0019	1.0008	0.0022	1.0019	0.0021
3.6	0.9998	0.0011	1.0002	0.0013	1.0009	0.0013	1.0016	0.0011
3.7	1.0000	0.0007	1.0003	0.0008	1.0008	0.0008	1.0012	0.0005
3.8	1.0001	0.0005	1.0003	0.0005	1.0006	0.0004	1.0009	0.0002
3.9	1.0001	0.0003	1.0003	0.0003	1.0005	0.0002	1.0006	0.0000
4.0	1.0001	0.0002	1.0002	0.0001	1.0003	0.0001	1.0004	-0.0001
4.1	1.0001	0.0001	1.0002	0.0001	1.0002	0.0000	1.0002	-0.0001
4.2	1.0001	0.0001	1.0001	0.0000	1.0001	0.0000	1.0001	-0.0001
4.3	1.0001	0.0000	1.0001	0.0000	1.0001	0.0000	1.0001	-0.0001
4.4	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000	-0.0001
4.5	—	—	—	—	—	—	1.0000	0.0000

TABLES OF $\operatorname{Re} \phi(z)$ AND $\operatorname{Im} \phi(z)$; $z = x + iy$

$y = 2.4$			$y = 2.5$	
x	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$	$\operatorname{Re} \phi(z)$	$\operatorname{Im} \phi(z)$
0.0	0.5000	3.7524	0.5000	4.5570
0.1	1.2027	3.6677	1.3971	4.4444
0.2	1.8586	3.4210	2.2301	4.1168
0.3	2.4258	3.0335	2.9414	3.6038
0.4	2.8710	2.5374	3.4882	2.9506
0.5	3.1734	1.9728	3.8417	2.2127
0.6	3.3253	1.3827	3.9940	1.4495
0.7	3.3322	0.8089	3.9552	0.7175
0.8	3.2115	0.2876	3.7515	0.0651
0.9	2.9888	-0.1533	3.4204	-0.4720
1.0	2.6955	-0.4965	3.0058	-0.8729
1.1	2.3638	-0.7351	2.5526	-1.1316
1.2	2.0245	-0.8722	2.1019	-1.2559
1.3	1.7031	-0.9188	1.6872	-1.2639
1.4	1.4193	-0.8916	1.3327	-1.1806
1.5	1.1852	-0.8100	1.0519	-1.0343
1.6	1.0062	-0.6938	0.8488	-0.8525
1.7	0.8815	-0.5615	0.7193	-0.6595
1.8	0.8058	-0.4280	0.6538	-0.4748
1.9	0.7708	-0.3046	0.6389	-0.3117
2.0	0.7670	-0.1987	0.6603	-0.1780
2.1	0.7845	-0.1136	0.7043	-0.0762
2.2	0.8147	-0.0501	0.7593	-0.0050
2.3	0.8504	-0.0062	0.8162	0.0394
2.4	0.8862	0.0209	0.8687	0.0624
2.5	0.9186	0.0349	0.9132	0.0697
2.6	0.9457	0.0395	0.9483	0.0667
2.7	0.9669	0.0380	0.9738	0.0577
2.8	0.9823	0.0330	0.9911	0.0463
2.9	0.9928	0.0266	1.0016	0.0347
3.0	0.9992	0.0201	1.0070	0.0243
3.1	1.0027	0.0142	1.0091	0.0159
3.2	1.0042	0.0095	1.0090	0.0095
3.3	1.0044	0.0059	1.0078	0.0051
3.4	1.0039	0.0033	1.0061	0.0022
3.5	1.0032	0.0016	1.0045	0.0005
3.6	1.0024	0.0006	1.0031	-0.0004
3.7	1.0017	0.0000	1.0020	-0.0008
3.8	1.0011	-0.0002	1.0012	-0.0008
3.9	1.0007	-0.0003	1.0006	-0.0007
4.0	1.0004	-0.0003	1.0003	-0.0006
4.1	1.0002	-0.0003	1.0001	-0.0004
4.2	1.0001	-0.0002	1.0000	-0.0003
4.3	1.0000	-0.0001	1.0000	-0.0002
4.4	1.0000	-0.0001	0.9999	-0.0001
4.5	1.0000	-0.0001	1.0000	-0.0001
4.6	1.0000	0.0000	1.0000	0.0000

REFERENCE

ROSSER, J. B. (1948): Theory and application of $\int_0^z e^{-x^2} dx$ and $\int_0^z e^{-p^2 y^2} dy \int_0^y e^{-x^2} dx$. Part I.

Methods of Computation, Mapleton House, Brooklyn, New York.

Paper received : September, 1959.

A FURTHER CONTRIBUTION TO THE NUMERICAL EVALUATION OF CERTAIN MULTIVARIATE NORMAL INTEGRALS

By PETER IHM

Botanic Institute, University of Freiburg, Germany

It has been shown in a previous paper (Ihm 1959) how the integral

$$I = \frac{1}{(2\pi)^{n/2} |\mathbf{A}|^{1/2}} \int_B e^{-\frac{1}{2} \mathbf{z}' \mathbf{A}^{-1} \mathbf{z}} dz_1 \dots dz_n \quad \text{with} \quad \mathbf{A} = \mathbf{D} + \frac{1}{c^2} \mathbf{i} \mathbf{i}'$$

can be solved numerically for real $\mathbf{i}$. $\mathbf{D}$ is a diagonal matrix, $\mathbf{i}$ a unit vector, B an n -dimensional interval; $\mathbf{A}$, $\mathbf{D}$ positively definite. It has also been shown that for

$$\mathbf{A}^* = b^2 \mathbf{I} + c^{-2} \mathbf{i} \mathbf{i}'$$

there exists an orthogonal matrix $\mathbf{S}$ giving

$$\mathbf{S}' \mathbf{A}^* \mathbf{S} = b^2 \mathbf{I} + \frac{1}{c^2} \begin{pmatrix} \mathbf{i}' \mathbf{i} & \mathbf{o} \\ \mathbf{o}' & \mathbf{O} \end{pmatrix}$$

where $\mathbf{o}$ and $\mathbf{O}$ are the null vector and the null matrix, respectively. For $\mathbf{i} = \mathbf{i}_1$ or $\mathbf{i} = i \mathbf{i}_1$ ($\mathbf{i}_1$ real) $\mathbf{A}^*$ corresponds to the class of all normal distributions having density constant on the surface of a hyperellipsoid which is rotationally symmetric about one axis. We have (see Ihm, 1959)

$$I = c(2\pi)^{-\frac{1}{2}(n+1)} |\mathbf{D}|^{-\frac{1}{2}} \int_{-\infty}^{\infty} \int_B \exp \left[-\frac{1}{2} c^2 \tau^2 - \frac{1}{2} (\mathbf{z} - \tau \mathbf{i})' \mathbf{D}^{-1} (\mathbf{z} - \tau \mathbf{i}) \right] dz_1 \dots dz_n d\tau.$$

with $\mathbf{i} = \mathbf{i}_1$. This holds also for $\mathbf{i} = i \mathbf{i}_1$. By repeated application of a theorem given by Cramér (1946, pp. 68-69) it can be shown that the sequence of integration may be inverted. Thus, for

$$I(\tau) = (2\pi)^{-\frac{1}{2}n} |\mathbf{D}|^{-\frac{1}{2}} \int_B \exp \left\{ -\frac{1}{2} (\mathbf{z} - \tau \mathbf{i})' \mathbf{D}^{-1} (\mathbf{z} - \tau \mathbf{i}) \right\} dz_1 \dots dz_n$$

$$I = c(2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} I(\tau) \exp \left(-\frac{1}{2} c^2 \tau^2 \right) d\tau.$$

The numerical evaluation for $\mathbf{i} = i \mathbf{i}_1$ is possible only if a table of the normal integral with complex argument is available. This function may be defined by

$$\phi(z) = \lim_{a \rightarrow \infty} \frac{1}{\sqrt{2\pi}} \int_{-a}^z e^{-\frac{1}{2} \zeta^2} d\zeta, \quad -\frac{\pi}{4} \leq \arg a \leq \frac{\pi}{4};$$

a table is published elsewhere in this journal. If the diagonal elements of $\mathbf{D}$ are σ_j^2 , and B is given by $z_j \leq a_j$ for $j = 1, 2, \dots, n$, $\mathbf{i}_1 = (e_j)$,

$$I(\tau) = \prod_{j=1}^n \phi(\{a_j - i\tau e_j\} / \sigma_j)$$

It may sometimes be more convenient to use $\tau^* = c\tau$ instead of τ . We then obtain

$$I = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} I(\tau^*/c) \exp(-\frac{1}{2}\tau^{*2}) d\tau^* = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \prod_{j=1}^n \phi(\{a_j - i\tau^* e_j / c\} / \sigma_j) \exp(-\frac{1}{2}\tau^{*2}) d\tau^*$$

REFERENCES

- CRAMÉR, H. (1946): *Mathematical Methods of Statistics*, Princeton University Press.
 IHM, P. (1959): Numerical evaluation of certain multivariate normal integrals. *Sankhyā*, **21**, 363-366.

Paper received : September, 1959.

ACKNOWLEDGEMENT

Thanks are due to the following for their kind help in the editing of the papers published in this issue.

Dr. R. R. Bahadur

Dr. D. Basu

Dr. P. K. Bhattacharyya

Shri S. John

Dr. C. R. Rao

Dr. J. Roy

Shri J. Sethuraman

Shri K. R. Shah

Editor,

Sankhyā : The Indian Journal of Statistics.

SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS

1. PLACE OF PUBLICATION : 204/1 Barrackpore Trunk Road, Calcutta-35
2. PERIODICITY OF ITS PUBLICATION : Quarterly
3. PRINTER'S NAME : Kalipada Mukherjee
NATIONALITY : Indian
ADDRESS : 204/1 Barrackpore Trunk Road, Calcutta-35
4. PUBLISHER'S NAME : Prafulla Chandra Mahalanobis
NATIONALITY : Indian
ADDRESS : 204/1 Barrackpore Trunk Road, Calcutta-35
5. EDITOR'S NAME : P. C. Mahalanobis
NATIONALITY : Indian
ADDRESS : 204 Barrackpore Trunk Road, Calcutta-35
6. NAMES AND ADDRESSES OF INDIVIDUALS
WHO OWN THE NEWSPAPER AND PART-
NERS OR SHAREHOLDERS HOLDING MORE
THAN ONE PER CENT OF THE TOTAL
CAPITAL : Published by Statistical Publishing Society
(non-profit distributing agency)

I, Prafulla Chandra Mahalanobis, hereby declare that the particulars given above are true to the best of my knowledge and belief.

15 May, 1961

Sd/ Prafulla Chandra Mahalanobis

BIOMETRIKA

Volume 48, Parts 1 and 2

June 1961

CONTENTS

Studies in the history of probability and statistics XI. Daniel Bernoulli on maximum likelihood.....	Kendall, M. G.
The variance of Spearman's rho in normal samples.....	David, F. N. and Mallows, C. L.
Tests for rank correlation coefficients. II.....	Feller, E. C. and Pearson, E. S.
Some methods of constructing exact tests.....	Durbin, J.
Preemptive priority queueing.....	Heathcote, C. R.
A two-sample sequential <i>t</i> -test.....	Hajnal, J.
Absolute and incomplete moments of the multivariate normal distribution.....	Nabeya, S.
Asymptotic expansions for the mean and variance of the serial correlation coefficient.....	White, John S.
Significance tests for paired-comparison experiments.....	Starks, T. H. and David, H. A.
Goodness-of-fit tests on a circle.....	Watson, G. S.
The use of orthogonal polynomials of the positive and negative binomial frequency functions in curve fitting by Aitken's method.....	Gonin, H. T.
The estimation of regression and error-scale parameters, when the joint distribution of the errors is of any continuous form and known apart from a scale parameter.....	Verhagen, A. M. W.
Latent vectors of random symmetric matrices.....	Mallows, C. L.
Expected values of normal order statistics.....	Harter, H. Leon
A distribution analogous to the Borel-Tanner.....	Haight, Frank A.
Occupancy probability distribution critical points.....	Nicholson, W. L.
Test of independence in intraclass 2x2 tables.....	Okamoto, Masashi and Ishii, Goro

Miscellanea

Reviews

Other Books Received

Corrigenda

The subscription, payable in advance, is 54/- (or \$8.00), per volume (including postage). Cheques should be made payable to *Biometrika*, crossed "A/c *Biometrika Trust*" and sent to the Secretary, *Biometrika* Office, Department of Statistics, University College, London, W. C. 1. All foreign cheques must be drawn on a Bank having a London agency.

Issued by THE BIOMETRIKA OFFICE, University College, London.

JOURNAL OF THE ROYAL STATISTICAL SOCIETY

Series B (Methodological)

40s each Part. Annual Subscription £3.2s (US \$9). Post free

CONTENTS OF VOL. 23, PART 2, 1961

A test of homogeneity of means under restricted alternatives (With Discussion).....	D. J. Bartholomew
Optimal programmes for varietal selection (With Discussion).....	R. N. Curnow
Some stochastic processes with absorbing barriers.....	T. L. Saaty
Some sequential tests using range.....	W. G. Gilchrist
First emptiness in a finite dam.....	B. Weesakul
The construction of optimal designs for the one-way classification analysis of variance.....	J. Aitchison
Some simple wear-dependent renewal processes.....	A. Mercer
An exact distribution of the Behrens-Fisher-Welch statistic for testing the difference between the means of two normal populations with unknown variances.....	W. D. Ray and A. E. N. T. Pitman
Closed queueing systems: A Generalisation of the machine interference model.....	F. Benson and G. Gregory
Testing for a jump in the spectral function.....	E. J. Hannan
Moments of a truncated bivariate normal distribution.....	S. Rosenbaum
On quadratic estimates of the interclass variance for unbalanced designs.....	R. S. Read
Mixtures of geometric distributions.....	H. E. Daniels
Prediction by exponentially weighted moving averages and related methods.....	D. R. Cox
The randomization bases of the problem of the amalgamation of weighted means.....	D. E. Barton and C. L. Mallows
Estimation from quantiles in destructive testing.....	H. P. Stout and F. Stern
A note on maximum-likelihood in the case of dependent random variables.....	S. D. Silvey
Some properties of the asymptotic variance of the sample quantiles and mid-ranges.....	P. K. Sen
Similarities between likelihoods and associated distributions <i>a posteriori</i>	D. A. Sprott
A method of maximum likelihood estimation.....	F. S. G. Richards
The intervals between regular events displaced in time by independent random deviations of large dispersal.....	T. Lewis
An approach to the scheduling of jobs on machines.....	E. Page

Obtainable through any bookseller or direct from the

ROYAL STATISTICAL SOCIETY

21 Bentinck Street, London, W. 1.

THE ANNALS OF MATHEMATICAL STATISTICS

(Founded by H. C. Carver)

The Official Journal of the Institute of Mathematical Statistics

CONTENTS OF VOLUME 32, NO. 2, JUNE, 1961

Georges Darmois, 1888-1960.....	D. Dugue
The existence and construction of balanced incomplete block designs.....	Haim Hanani
Random allocation designs II: Approximate theory for simple random allocation.....	A. P. Dempster
Sampling moments of means from finite multivariate populations.....	D. W. Behnken
On the foundations of statistical inference: Binary experiments.....	Allan Birnbaum
Some extensions of the idea of bias.....	H. R. van der Vaart
Multivariate correlation models with mixed discrete and continuous variables.....	I. Olkin and R. F. Tate
Limits for a variance component with an exact confidence coefficient.....	W. C. Healy, Jr.
Confidence sets for multivariate medians.....	P. G. Hoel and E. M. Scheuer
Distribution free tests of independence based on the sample distribution function.....	J. R. Blum, J. Kiefer, and M. Rosenblatt
Some exact results for one-sided distribution tests of the Kolmogorov-Smirnov type.....	P. Whittle
Some extensions of the Wald-Wolfowitz-Noether theorem.....	Jaroslav Hajek
The gap test for random sequences.....	Eve Bofinger and V. J. Bofinger
The multivariate saddlepoint method and chi-squared for the multinomial distribution.....	I. J. Good
A generalization of Wald's identity with applications to random walks.....	H. D. Miller
A characterization of the weak convergence of measures.....	Robert Bartoszyński
Exponential bounds on the probability of error for a discrete memoryless channel.....	Samuel Kotz
An exponential bound on the strong law of large numbers for linear stochastic processes with absolutely convergent coefficients.....	L. H. Koopmans
Expected utility for queues servicing messages with exponentially decaying utility.....	Frank A. Haight
On the coding theorem for the noiseless channel.....	Patrick Billingsley

NOTES

The essential completeness of the class of generalized sequential probability ratio tests.....	M. H. De Groot
A problem in survival.....	James B. MacQueen
First passage time for a particular Gaussian process.....	D. Slepian
A note on the Ergodic theorem of information theory.....	K. L. Chung
Remark concerning two-state semi-Markov processes.....	Cyrus Derman
An example of an ancillary statistic and the combination of two samples by Bayes' theorem.....	D. A. Sprott

Correction Notes
Abstracts of Papers
News and Notices
Publications Received

Subscription Rate: \$ 10.00 per year

Address orders for subscriptions and back numbers to the
Secretary, Institute of Mathematical Statistics,
Department of Statistics, University of
North Carolina, Chapel Hill, North Carolina.

ECONOMETRICA

Vol. 29, No. 2

April, 1961

CONTENTS

Aggregation of variables in dynamic systems.....	Herbert A. Simon and Albert Ando
On the cost of approximate specification in simultaneous equation estimation.....	Franklin M. Fisher
The graduation of income distributions.....	Peter R. Fisk
Behavioristic foundations of utility.....	Patrick Suppes
Nonlinear programming by the simplex method.....	H. O. Hartley
Aggregation in Leontief matrices and the labour theory of value.....	M. Morishima and F. Seton
A note on "Aggregation in Leontief matrices and the labour theory of value".....	Leif Johansen
A note on the residual variance estimation in simultaneous equations.....	A. L. Nagar
A note on the general possibility theorem of the social welfare function.....	Yasusuke Murakami
On the existence of general equilibrium: Some corrections.....	Lionel W. McKenzie

Report of the St. Louis Meeting
Report of the Stanford Meeting
Book Reviews
Letter to the Editor
Announcements and Notes

*Subscription price for nonmembers: \$ 12.00 per year in the United States and Canada;
\$ 10.00 per year in all other Countries. Single copies \$ 3.00*
Graduate Students recommended by a faculty member: \$ 4.00 per year.

Subscriptions should be sent to

**The Secretary
ECONOMETRIC SOCIETY**

Box 1264, Yale Station, New Haven, Connecticut, U.S.A.

SANKHYĀ

THE INDIAN JOURNAL OF STATISTICS

Edited by : P. C. MAHALANOBIS

SERIES A, VOL. 23

AUGUST 1961

PART 3

AN ELEMENTARY THEORY OF THE BROWNIAN MOTION RANDOM FUNCTION*

By J. KAMPÉ DE FÉRIET

University of Lille, France

SUMMARY. By making a systematic use of the existence of a Schauder basis in the Banach space $C_0[0, 1]$ consisting of all continuous functions defined on the interval $[0, 1]$ and vanishing at 0 and 1, the Brownian motion random function is expanded into an infinite series which is *almost surely, absolutely and uniformly* convergent over every bounded interval. This shows that the Brownian motion random function is continuous with probability one. The same idea is used to define the well-known Wiener integral for square integrable functions. The importance of this method lies in the fact that it leads to an extremely elementary approach to the theory of Brownian motion.

1. INTRODUCTION

In his pioneering work, N. Wiener (1923) for the first time has given a rational theory of the random function $W(t)$, representing the abscissa, at time t , of a particle, starting at time 0 from the origin and submitted to the random impulsions of a liquid in Brownian motion. But his point of view (the more complete expositions are to be found in Wiener (1930)), Paley and Wiener (1941) and based on a definition of a measure in the space of all real functions—is highly abstract. A more elementary approach has been given by Paul Lévy (1948) and (1952) consisting essentially in an approximation of the curve $x = W(t)$ by a sequence of polygonal lines.

In our researches on probability measures in Banach spaces (Férier (1957) and 1960)), we have noticed that using the triangular functions $e_n(t)$ (which define a basis in the Banach space $C_0[0, 1]$ we were (as a very particular case of our general results) able to prove that

$$W(t) - tW(1) = \sum_{n=1}^{+\infty} \eta_n e_n(t) \quad \dots \quad (1.1)$$

*A lecture given at the Indian Statistical Institute, Calcutta; notes prepared by K. R. Parthasarathy, R. Ranga Rao, and S. R. S. Varadhan.

(where $\eta_1, \eta_2, \dots$ are normal independent random variables, independent of $W(1)$), the series being almost surely, absolutely and uniformly convergent in $[0, 1]$.

The sums

$$x_n = \sum_1^n \eta_k e_k(t) \quad \dots \quad (1.2)$$

correspond of course to polygonal lines and, from this point of view, we are on the same line of approach as that of Paul Lévy. But the systematic use of the triangular functions $e_n(t)$ gives us a formalization of his ideas, which seems to us to clarify and simplify the problem. The series (1.1) is to be found in a casual remark in our papers (Féret (1957), p. 816 and (1960), p. 153) on probability measures in $C_0[0, 1]$.

Our proof of the almost sure, absolute and uniform convergence of (1.1) given by the author (Féret, (1957)) is quite elementary and can be used in any course on probability for Graduate Students. Thus it seems worth while to sketch an exposition of the theory of the Brownian motion function $W(t)$ based on this simple representation; we hope, that this small monograph could be useful for the research workers in Theoretical Physics, who use $W(t)$, but are not always familiar with measures in function spaces.

2. PRELIMINARIES

Let $J_1 = [0, 1]$ be the semi-open unit interval, and N_q , the set of all integers n such that $2^{q-1} \leq n < 2^q$. For any integer n there exists a unique pair of integers

(q_n, p_n) such that $n = 2^{q_n-1} + p_n$ where $0 \leq p_n \leq 2^{q_n-1} - 1$.

Let

$$J_n = \left\{ t : \frac{p_n}{2^{q_n-1}} \leq t < \frac{p_n+1}{2^{q_n-1}} \right\} \quad \dots \quad (2.1)$$

$$t_n = \frac{2p_n + 1}{2^{q_n}} \quad \dots \quad (2.2)$$

Thus, for instance, $J_1 = [0, 1]$, $J_2 = [0, \frac{1}{2}]$, $J_3 = [\frac{1}{2}, 1]$ and so on. As n varies over N_q , the class of sets $\{J_n\}$ defines a partition of J_1 , i.e.,

$$J_m \cap J_n = \phi \quad \text{if } m \neq n$$

$$\bigcup_{n \in N_q} J_n = J_1.$$

For any $t \in J_1$, it is clear that there exists a unique sequence $\{n_q\}$ such that

$$(1.1) \quad \dots \quad t = \bigcap_{q=1}^{\infty} J_{n_q}.$$

The point t_n defined by (2.2) is the midpoint of the interval J_n and as n varies over the integers $n = 1, 2, \dots$, t_n runs over all binary points of J_1 .

AN ELEMENTARY THEORY OF BROWNIAN MOTION RANDOM FUNCTION

Now we consider the space $C_0[0, 1]$ of real valued continuous functions defined on $[0, 1]$ and vanishing at 0 and 1.

For any $x(t) \in C_0[0, 1]$, we write

$$\|x\| = \sup_{0 \leq t \leq 1} |x(t)|. \quad \dots (2.3)$$

Under the norm (2.3) the space $C_0[0, 1]$ becomes a separable Banach space. The essential idea of the present approach is to exploit the existence of a very simple basis in the space $C_0[0, 1]$. By a basis in a Banach space $\mathcal{X}$ we mean a sequence x_n of elements in $\mathcal{X}$ such that every element $x \in \mathcal{X}$ has a unique representation of the form

$$x = \sum_{n=1}^{\infty} \eta_n x_n.$$

A basis for the space $C_0[0, 1]$ in the above sense was given by Schauder (1927) with the help of functions $e_n(t)$, ($n = 1, 2, \dots$) defined as follows,

$$e_n(t) = \begin{cases} 0 & \text{for } t \in J_n \\ 2^{q_n} \left(t - \frac{p_n}{2^{q_n-1}} \right) & \text{for } t \in J_{2n} \\ 2^{q_n} \left(\frac{p_n+1}{2^{q_n-1}} - t \right) & \text{for } t \in J_{2n+1} \end{cases} \quad \dots (2.4)$$

Since $J_{2n} \cap J_{2n+1} = \phi$ and $J_{2n} \cup J_{2n+1} = J_n$, the above function is well defined. As an aid to understanding it may be added that the function is the triangular function vanishing outside J_n and taking its maximum value 1 at the midpoint t_n of J_n .

Theorem 1 (Schauder): For any $x(t) \in C_0[0, 1]$ there exists a unique sequence $\eta_1, \eta_2, \dots, \eta_n, \dots$ of real numbers such that the infinite series $\sum_{r=1}^{\infty} \eta_r e_r(t)$ converges uniformly to $x(t)$.

Proof: Let us write

$$\begin{aligned} x_1(t) &= \eta_1 e_1(t) \\ x_2(t) &= \eta_1 e_1(t) + \eta_2 e_2(t) \\ &\dots \dots \dots \\ x_n(t) &= \eta_1 e_1(t) + \dots + \eta_n e_n(t) = x_{n-1}(t) + \eta_n e_n(t) \end{aligned}$$

where the constants $\eta_1, \eta_2, \dots, \eta_n$ are determined by the requirement that $x_j(t)$ agrees with $x(t)$ at the points $t_1, t_2, \dots, t_j$ (the t_j 's defined by (2.2)). More precisely,

$$\begin{aligned} \eta_1 &= x(t_1) \\ \eta_2 &= x(t_2) - x_1(t_2) \\ &\dots \dots \dots \\ \eta_n &= x(t_n) - x_{n-1}(t_n). \end{aligned}$$

An elementary computation shows that

$$\eta_n = x \left(\frac{2p_n+1}{2q_n} \right) - \frac{1}{2} \left[x \left(\frac{p_n}{2q_n-1} \right) + x \left(\frac{p_n+1}{2q_n-1} \right) \right]. \quad \dots (2.5)$$

It is now evident that the function $x_n(t)$ constructed in this way is merely a polygonal line P_n agreeing with $x(t)$ at the points $t_1, t_2, \dots, t_n$ and is linear in between. It is then obvious that $x_n(t)$ converges uniformly to $x(t)$ in J_1 . The polygonal line P_n agrees with P_{n-1} except on J_n , where the side AB of P_{n-1} is replaced by a triangle ABC , C having t_n as abscissa and $x(t_n)$ as ordinate. This property is an important feature of the approximation of $x(t)$ by the sequence $x_n(t)$: a new step in the approximation does not alter the result previously obtained; the fitting of a new point $x(t_n) = x_n(t_n)$ does not destroy the preceding fittings.

To see uniqueness suppose

$$x(t) = \sum_1^{\infty} \eta'_n e_n(t)$$

is another representation. Since $e_n(t_j) = 0$ for all $n > j$, this leads to the same relations as those obtained for the definition of η_n in the last paragraph and consequently $\eta'_n = \eta_n$ where η_n is given by (2.5).

Thus the sequence of functions $e_n(t)$ constitute a basis for $C_0[0, 1]$, the Schauder basis. The expansion of $x(t)$ in terms of $e_1, e_2, \dots$, is called the Schauder series of $x(t)$.

The Schauder series for $x(t)$ converges uniformly in t . However, in general, there is no absolute convergence. A sufficient condition for this to happen is given by the following theorem.

Theorem 2 : *The series $\sum_1^{\infty} \eta_n e_n(t)$ converges absolutely and uniformly to $[0, 1]$ whenever*

$$\sum_{q=1}^{\infty} \sup_{n \in N_q} |\eta_n| < +\infty. \quad \dots (2.6)$$

Proof : Since $\{J_n\}$, $n \in N_q$ is a partition of $[0, 1]$ only one term in $\sum_{n \in N_q} |\eta_n| e_n(t)$ can be non-zero. Thus

$$\sum_{n \in N_q} |\eta_n| e_n(t) \leq \sup_{n \in N_q} |\eta_n|.$$

The theorem is an immediate consequence of this inequality. Let us note that

$$|x(t)| \leq \sum |\eta_n| e_n(t) \leq \sum_{q=1}^{\infty} \sup_{n \in N_q} |\eta_n|. \quad \dots (2.7)$$

AN ELEMENTARY THEORY OF BROWNIAN MOTION RANDOM FUNCTION

Let us now recall the definition of the orthonormal basis (Haar, 1909) in $L^2[0, 1]$. This is defined by the following sequence of functions $h_n(t)$:

$$h_0(t) = 1$$

$$h_n(t) = \begin{cases} 0 & \text{if } t \in J_n \\ \frac{q_n-1}{2} & \text{if } t \in J_{2n} \\ -\frac{q_n-1}{2} & \text{if } t \in J_{2n+1} \end{cases} \quad \dots (2.8)$$

It is clear that

$$e_n(t) = 2^{\frac{q_n+1}{2}} \int_0^t h_n(s) ds, \quad n = 1, 2, \dots \quad \dots (2.9)$$

In the following pages, as far as the theory of probability is concerned we will make use of the two following known criteria.

Theorem 3: *Let $[X_1, X_2, \dots, X_n, \dots]$ be a sequence of random variables. Then*

$$\left[\sum_1^\infty E|X_n| < \infty \right] \implies \text{prob} \left[\sum_1^\infty |X_n| < \infty \right] = 1.$$

This theorem is a simple consequence of Fatou's lemma.

We note here that no hypothesis is made concerning the independence of the random variables $X_1, X_2, \dots, X_n, \dots$.

Theorem 4: *The n random variables $[X_1, \dots, X_n]$ follow an n -variate normal law if and only if, for any constants $A_1, \dots, A_n$ the linear combination*

$$X = A_1 X_1 + \dots + A_n X_n$$

is normally distributed.

3. THE WIENER-LEVY RANDOM FUNCTION

Let $\eta_1, \dots, \eta_n, \dots$ be a sequence of mutually independent and normally distributed random variables such that

$$E\eta_n = 0, \quad E\eta_n^2 = \frac{\sigma^2}{2^{q_n+1}}, \quad E\eta_m \eta_n = 0 \quad \text{for } m \neq n. \quad \dots (3.1)$$

Theorem 5: *The series*

$$w(t) = \sum_1^\infty \eta_n e_n(t) \quad \dots (3.2)$$

is almost surely, absolutely and uniformly convergent in $[0, 1]$. Its sum $w(t)$ is a normal random function continuous with probability one. The mean and covariance of $w(t)$ are given by

$$E w(t) = 0$$

$$\begin{aligned} \Gamma(t, s) = E w(t) w(s) &= \sigma^2 t(1-s) \quad 0 \leq t \leq s \leq 1 \\ &= \sigma^2 s(1-t) \quad 0 \leq s \leq t \leq 1. \end{aligned}$$

Proof: From Hölder's inequality it follows that

$$E[\sup_{n \in N_q} |\eta_n|] \leq [E(\sup_{n \in N_q} |\eta_n|)^4]^{1/4} = [E \sup_{n \in N_q} |\eta_n|^4]^{1/4}.$$

But obviously

$$E[\sup_{n \in N_q} |\eta_n|^4] \leq \sum_{n \in N_q} E |\eta_n|^4$$

and from (3.1)

$$E[\sup_{n \in N_q} |\eta_n|^4] \leq 3\sigma^4 \sum_{n \in N_q} \frac{1}{2^{2q_n+2}} = \frac{3\sigma^4}{2^{q+3}}$$

and thence

$$E[\sup_{n \in N_q} |\eta_n|] \leq \frac{3^{1/4} \sigma}{2^{q/4}}.$$

Thus we have

$$\sum_{q=1}^{\infty} E \sup_{n \in N_q} |\eta_n| \leq 3^{1/4} \sigma \sum_{q=1}^{\infty} 2^{-q/4} < \infty.$$

Consequently, by Theorems 2 and 3 the series (3.2) converges absolutely and uniformly with probability one. At the same time this proves that $w(t)$ is almost surely continuous in $[0, 1]$.

Now, to prove that $w(t)$ is a normal random function we have to show that for any finite set $[\tau_1, \dots, \tau_m]$ of values of $t \in [0, 1]$ the n random variables $w(\tau_1), \dots, w(\tau_m)$ follow a m -variate normal law. Let us consider the linear combination:

$$X = A_1 w(\tau_1) + \dots + A_m w(\tau_m).$$

We have obviously

$$X = \sum_1^{+\infty} \eta_n [A_1 e_n(\tau_1) + \dots + A_m e_n(\tau_m)]. \quad \dots (3.3)$$

But
$$\sum_{n \in N_q} |\eta_n| |A_1 e_n(\tau_1) + \dots + A_m e_n(\tau_m)| \leq [|A_1| + \dots + |A_m|] \sup_{n \in N_q} |\eta_n|.$$

Thus (applying again Theorems 2 and 3) the series (3.3) is (even absolutely) convergent with probability one; X being the sum of a convergent series of normal independent random variables, is normally distributed and the conclusion follows by Theorem 4.

AN ELEMENTARY THEORY OF BROWNIAN MOTION RANDOM FUNCTION

The n -variate normal law being completely determined by the values of the mean $E[w(t)]$ and the covariance $E[w(t)w(s)]$, let us compute these functions.

First, from (3.1) we have :

$$E[w(t)] = 0$$

(by term by term integration in the probability space of $\eta_1, \dots, \eta_n, \dots$) and

$$(3.4) \quad \Gamma(t, s) = Ew(t)w(s) = \sum_{m,n} E \eta_m \eta_n e_m(t) e_n(s) = \sum_{n=1}^{\infty} \frac{\sigma^2}{2^{q_n+1}} e_n(t) e_n(s). \dots (3.5)$$

We shall prove that $\Gamma(t, s)$ is the required function by comparing it with the Green function of the following problem :

$$(3.6) \quad y''(t) = -x(t), \quad y(0) = y(1) = 0. \dots (3.6)$$

From (3.5) (due to the uniform convergence of (3.2))

$$(3.7) \quad \begin{aligned} \int_0^1 \Gamma(t, s) x(s) ds &= \sum_1^{\infty} \frac{\sigma^2}{2^{q_n+1}} e_n(t) \int_0^1 e_n(s) x(s) ds \\ &= - \sum_1^{\infty} \frac{\sigma^2}{2^{q_n+1}} e_n(t) \int_0^1 e_n(s) y''(s) ds. \end{aligned}$$

But integrating by parts and using (2.9),

$$(3.8) \quad \begin{aligned} - \int_0^1 e_n(s) y''(s) ds &= 2^{\frac{q_n+1}{2}} \int_0^1 h_n(s) y'(s) ds \\ &= 2^{\frac{q_n+1}{2}} \left[2^{\frac{q_n-1}{2}} \int_{2^n}^1 y'(s) ds - 2^{\frac{q_n-1}{2}} \int_0^{2^{n+1}} y'(s) ds \right] \\ &= 2^{q_n+1} \left\{ y \left(\frac{2^{q_n+1}}{2^{q_n}} \right) - \frac{1}{2} \left[y \left(\frac{2^{q_n}}{2^{q_n-1}} \right) + y \left(\frac{2^{q_n+1}}{2^{q_n-1}} \right) \right] \right\} \\ &= 2^{q_n+1} b_n, \text{ say.} \end{aligned}$$

$$(3.9) \quad \text{Thus} \quad \int_0^1 \Gamma(t, s) x(s) ds = \sigma^2 \sum b_n e_n(t). \dots (3.7)$$

But interpreting the values of the coefficients b_n through (2.5) we see that the right hand side of (3.7) is simply the Schauder series of $y(t)$ and

$$(3.10) \quad \int_0^1 \Gamma(t, s) x(s) ds = \sigma^2 y(t)$$

which shows that $\Gamma(t, s)$ is effectively (up to the factor σ^2) the Green function of (3.6); this Green function is $t(1-s)$ in $0 \leq t \leq s \leq 1$ and $s(1-t)$ in $0 \leq s \leq t \leq 1$. The proof of the theorem is complete.

So far we restricted ourselves to the description of the normal random function $w(t)$ in the interval $[0, 1]$. Now we will define a normal random function $W(t)$ on $[0, \infty]$. To this purpose, we introduce

$$e_0(t) = \begin{cases} 0 & \text{for } t \leq 0 \\ t & \text{for } 0 \leq t \leq 1 \\ 1 & \text{for } t > 1 \end{cases} \quad \dots (3.8)$$

and write

$$W_j(t) = \xi_j e_0(t-j+1) + \sum_1^{\infty} \eta_{jn} e_n(t-j+1), \quad j = 1, 2, \dots \quad \dots (3.9)$$

where $e_1, e_2, \dots$ are the functions introduced earlier (cf. 2.4) their definition being extended to $(-\infty, +\infty)$ by an extension of the convention

$$e_n(t) = 0, \quad t \in J_n. \quad \dots (3.10)$$

and

$$\xi_1, \xi_2, \dots; \eta_{j1}, \eta_{j2}, \dots, \eta_{jn} \quad j = 1, 2, \dots$$

are independent normal random variables with

$$\begin{aligned} E \xi_j &= 0, & E \xi_j^2 &= \sigma^2, \\ E \eta_{jn} &= 0, & E \eta_{jn}^2 &= \frac{\sigma^2}{2^{qn+1}}. \end{aligned} \quad \dots (3.11)$$

Let us define

$$W(t) = \sum_{j=1}^{\infty} W_j(t). \quad \dots (3.12)$$

Then, for any positive integer m we have in the interval $m-1 \leq t < m$,

$$W(t) = \xi_1 + \dots + \xi_{m-1} + \xi_m e_0(t-m+1) + \sum_{n=1}^{\infty} \eta_{mn} e_n(t-m+1). \quad \dots (3.13)$$

From the same reasoning as in the proof of Theorem 5 we see that the series is almost surely, absolutely and uniformly convergent in $[m-1, m]$. Thus, applying this result successively to the intervals $[0, 1], [1, 2], \dots, [m-1, m]$ we see that $W(t)$, defined by (3.12), is a normal random function, continuous with probability one in any finite interval.

Now let us put

$$w_m(t) = \sum_1^{+\infty} \eta_{mn} e_n(t-m+1).$$

AN ELEMENTARY THEORY OF BROWNIAN MOTION RANDOM FUNCTION

The computations of $E[w(t)]$ and $E[w(t)w(s)]$ made in $[0, 1]$ apply without any modification to $w_m(t)$ in $[m-1, m]$; replacing 0 by $m-1$ and 1 by m .

$$E[w_m(t)] = 0$$

$$\begin{aligned} E[w_m(t) w_m(s)] &= \sigma^2(t-m+1)(m-s), & m-1 \leq t \leq s \leq m \\ &= \sigma^2(s-m+1)(m-t), & m-1 \leq s \leq t \leq m. \end{aligned}$$

Thus, from (3.13), we have

$$E[W(t)] = 0 \quad \dots (3.14)$$

$$\begin{aligned} E[W(t)W(s)] &= \sigma^2(m-1) + \sigma^2(t-m+1)(s-m+1) + \sigma^2(t-m+1)(m-s) \\ &= \sigma^2 t, & m-1 \leq t \leq s \leq m \end{aligned} \quad \dots (3.15)$$

$$\begin{aligned} \text{and } E[W(t)W(s)] &= \sigma^2(m-1) + \sigma^2(t-m+1)(s-m+1) + \sigma^2(s-m+1)(m-t) \\ &= \sigma^2 s, & m-1 \leq s \leq t \leq m. \end{aligned} \quad \dots (3.16)$$

We have supposed that both t and s are in the same interval $[m-1, m]$; if

$$p-1 \leq t < p$$

$$q-1 \leq s < q$$

the computations are even simpler; from (3.13) we have

$$W(t) = \xi_1 + \dots + \xi_{p-1} + \xi_p e_0(t-p+1) + \sum_1^\infty \eta_{pn} e_n(t-p+1)$$

$$W(s) = \xi_1 + \dots + \xi_{q-1} + \xi_q e_0(s-q+1) + \sum_1^\infty \eta_{qn} e_n(s-q+1).$$

Thus, if $p \leq q-1$

$$E[W(t)W(s)] = \sigma^2(p-1) + \sigma^2(t-p+1) = \sigma^2 t \quad \dots (3.17)$$

and if $q \leq p-1$

$$E[W(t)W(s)] = \sigma^2(q-1) + \sigma^2(s-q+1) = \sigma^2 s. \quad \dots (3.18)$$

Thus from (3.15), (3.16), (3.17) and (3.18) we conclude that for any finite values of t and s

$$E[W(t)W(s)] = \sigma^2 \min(t, s). \quad \dots (3.19)$$

Now we come to the crucial point; the mean and covariance determine completely the separable version (see, for this concept J. L. Doob (1953)p. 392) of a normal random function. The Brownian motion function of N. Wiener-P. Lévy is by definition the separable version of the normal random function having

$$\text{mean} = 0$$

$$\text{covariance} = \sigma^2 \min(t, s).$$

Theorem 6 : *The normal random function $w(t)$ defined in any finite interval by the almost surely, absolutely and uniformly convergent series (3.12) is identical with the Wiener-Lévy function.*

Let us point out that from (3.14) and (3.19) follows immediately the fundamental property of $W(t)$, which very often is taken as its characteristic feature.

Corollary : For any finite set :

$$0 \leq t_1 < t_2 \leq t_3 < t_4 \leq \dots \leq t_{2n-1} < t_{2n} < +\infty$$

the n random variables

$$U_j = W(t_{2j}) - W(t_{2j-1}), \quad j = 1, \dots, n \quad \dots \quad (3.20)$$

are independent, normally distributed and

$$E(U_j) = 0 \quad \dots \quad (3.21)$$

$$E(U_j^2) = \sigma^2(t_{2j} - t_{2j-1}). \quad \dots \quad (3.22)$$

We will now prove the following theorem.

Theorem 7 : *The normal random function $W(t)$ is almost surely not of bounded variation on any given interval.*

Proof: We will restrict ourselves to the case when the interval is $[0, 1]$. The increment of $W(t)$ on J_n , for $n \in N_q$,

$$\delta q_n = W\left(\frac{p_n+1}{2^{q_n-1}}\right) - W\left(\frac{p_n}{2^{q_n-1}}\right)$$

is a normal random variable and

$$E\delta q_n = 0, \quad E\delta q_n^2 = \frac{\sigma^2}{2^{q_n-1}}. \quad \dots \quad (3.23)$$

Moreover, the 2^{q-1} increments δq_n when n runs through N_q are independent random variables, because the $J_n, n \in N_q$ define a binary partition of J_1 . Thus

$$V_q = \sum_{n \in N_q} |\delta q_n| \quad \dots \quad (3.24)$$

defines the variation of $W(t)$ over the binary partition of order q by a sum of 2^{q-1} independent random variables. Since

$$J_n = J_{2n} \cup J_{2n+1}$$

and

$$2n, 2n+1 \in N_{q+1}$$

one has

$$\delta q_n = \delta q_{2n} + \delta q_{2n+1}.$$

Thus

$$V_q \leq V_{q+1}. \quad \dots \quad (3.25)$$

AN ELEMENTARY THEORY OF BROWNIAN MOTION RANDOM FUNCTION

To establish the theorem we will prove that

$$\text{prob} [\lim_{q \rightarrow \infty} V_q < +\infty] = 0 \quad \dots (3.26)$$

$$EV_q = \sum_{n \in N_q} E|\delta q_n| = \sum_{n \in N_q} \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma}{2^{\frac{q_n-1}{2}}} = 2^{\frac{q-1}{2}} \sqrt{\frac{2}{\pi}}.$$

Thus $\lim_{q \rightarrow \infty} EV_q = \infty$. On the other hand the variance of V_q satisfies the following inequality

$$\text{var} (V_q) = \sum \text{var} (|\delta q_n|) \leq \sum E|\delta q_n|^2 = \sigma^2.$$

Now consider the random variable

$$Z_q = \frac{V_q - EV_q}{\sqrt{EV_q}}.$$

Note that

$$EZ_q = 0$$

and

$$EZ_q^2 \leq \frac{\sigma^2}{EV_q}.$$

Hence $Z_q \rightarrow 0$ in probability which together with (3.25) implies that $V_q \rightarrow \infty$ with probability one.*

4. INTEGRATION WITH RESPECT TO WIENER-LÉVY RANDOM FUNCTION

The problem of integration with respect to Wiener-Lévy random function is to give a meaningful definition of the integral

$$\int f(t) dW(t). \quad \dots (4.1)$$

At the outset it is clear that the integral (4.1) has no meaning in the classical Lebesgue-Stieltjes sense, since $W(t)$ is almost never a function of bounded variation. A method of defining the integral is given by Doob (1953) as follows
First, one defines the integrals

$$\int_a^b dW(t)$$

as $W(b) - W(a)$. The definition is then extended to step-functions by linearity. A completion procedure then leads to the definition of the integral for all square integrable functions. The completion procedure here is in the sense of mean convergence. The elementary approach to the Brownian motion process given earlier at once

* This proof, shorter than the one given in my lecture, is due to R. Ranga Rao.

leads to a definition in which the integral is obtained as the almost sure limit of a certain well-defined sequence of random variables.

Let $W(t)$ be the Wiener-Lévy random function and $F(t)$ be any square integrable function in the interval $[0, \infty]$. By definition (see (3.1) and (3.2))

$$W(t) = \sum_{j=1}^{\infty} W_j(t)$$

where
$$W_j(t) = \xi_1 + \dots + \xi_{j-1} + \xi_j e_0(t-j+1) + \sum_{n=1}^{\infty} \xi_{nj} e_n(t-j+1),$$

$(\xi_1, \dots, \xi_n \dots)$, η_{nj} being mutually independent random variables satisfying (3.11).

Now, in order to define $\int_0^{\infty} F(t) dW(t)$ it is enough to define it for the interval

$[0, 1]$. For then it is possible to define the integral over $[m-1, m)$ for any integer m and then extend the definition by writing

$$\int_0^{\infty} F(t) dW(t) = \sum_{m=1}^{\infty} \int_{m-1}^m F(t) dW(t) = \sum_{m=1}^{\infty} \int_{m-1}^m F(t) dW_m(t).$$

We shall, therefore, confine our attention to the interval $[0, 1]$. In this interval

$$W(t) = t \xi_1 + \sum_{n=1}^{\infty} \eta_{n1} e_n(t).$$

Since each $e_n(t)$ is a function of bounded variation,

$$\int_0^1 F(t) de_n(t) = 2^{(q_n+1)/2} \int_0^1 F(t) h_n(t) dt = 2^{(q_n+1)/2} F_n,$$

where

$$F_n = \int_0^1 h_n(t) F(t) dt.$$

are simply the Fourier coefficients of $F(t)$ with respect to the Haar functions. Since the function $F \in L^2[0, 1]$ it follows that

$$\sum_0^{+\infty} F_n^2 = \int_0^1 [F(t)]^2 dt < \infty.$$

AN ELEMENTARY THEORY OF BROWNIAN MOTION RANDOM FUNCTION

The definition of Wiener-integral may then be given as follows.

Theorem 8: Let the pseudo-Stieltjes integral be defined as follows

$$\int_0^1 F(t) dW(t) = \xi_1 F_0 + \sum_{n=1}^{\infty} \eta_n 2^{(q_n+1)/2} F_n. \quad \dots (4.2)$$

Then the integral is well defined (i.e. the series on the right converges with probability one). Moreover, if X_F denotes the integral (4.2), then

$$EX_F = 0, EX_F^2 = \int_0^1 [F(t)]^2 dt \quad \dots (4.3)$$

and for any two $F, G \in L^2[0, 1]$

$$E(X_F X_G) = \int_0^1 F(t)G(t)dt. \quad \dots (4.4)$$

The correspondence between the Hilbert spaces of all functions $F(t) \in L^2[0, 1]$ and of all normal random variables X such that $EX = 0, EX^2 < \infty$ is thus an isometric isomorphism.

For a proof we have only to verify the convergence of the series

$$\sum_{n=1}^{\infty} \eta_{n1} 2^{(q_n+1/2)} F_n.$$

Now let

$$y_n = \eta_{n1} 2^{(q_n+1/2)} F_n.$$

Then clearly

$$Ey_n = 0, Ey_n^2 = E\eta_{n1}^2 2^{q_n+1} F_n^2 = \sigma^2 F_n^2.$$

Also, the random variables y_n are mutually independent and therefore, by Kolmogorov's theorem, the series converges with probability one.

If $X_F = \int F(t) dW(t) = \xi_1 F_0 + \sum \eta_{n1} 2^{(q_n+1)/2} F_n$ and $X_G = \xi_1 G_0 + \sum \eta_{n1} 2^{(q_n+1)/2} G_n$,

it is easily verified that

$$E(X_F X_G) = \sum_0^{\infty} F_n G_n = \int_0^1 F(t)G(t)dt.$$

This demonstrates the theorem completely.

REFERENCES

- DOOB, J. L. (1953): *Stochastic Processes*, John Wiley, New York.
- HAAR, A. (1909): Zur Theorie der orthogonalen Funktionen Systeme. Inaug. Dissert., Gottingen.
- FÉRIET, J. KAMPÉ DE (1957): Mesures de probabilité sur l'espace de Banach $C[0, 1]$. *Comptes Rendus*, Paris, **240**, 813-816.
- (1960): Mesures de probabilité sur les espaces de Banach possédant une base dénombrable. *J. Math. Pures. Appl.*, **39**, 97-163.
- LÉVY, P. (1948): *Processus Stochastiques et Mouvement Brownien* Gauthier Villaras, Paris.
- (1952): Le mouvement Brownien. *Mem. Sciences Math.*, 126.
- PALEY, R. E. A. C. and WIENER N. (1934): Fourier transforms in the complex domain. *Am. Math. Soc. Colloquium*, XXI. New York.
- SCHAUDER, J. (1927): Zur theorie stetiger abbildungen in funktionalraum. *Math. Zeitschrift.*, **26**, 47-65.
- WIENER, N. (1923): Differential space. *J. Math. Phy.*, M. I. T., **2**, 131-174.
- (1930): Generalised harmonic analysis. *Acta. Math.*, **55**, 117-258.

Paper received : June, 1961.

ON TESTING THE MEAN OF A DISCRETE LINEAR PROCESS

By K. R. PARTHASARATHY

Indian Statistical Institute

SUMMARY. In this paper a method of testing the mean function of a discrete linear stationary process is given when the spectral density function is known. The asymptotic distribution of the proposed test statistic is derived.

1. INTRODUCTION

The problem of testing the mean function of a linear process has been solved by Grenander and Rosenblatt (1951) in the case when the process is Gaussian and a simple hypothesis is being tested against a simple alternative. In this paper we consider the problem of testing the null hypothesis that the mean function of an arbitrary linear process is zero against all alternatives. The asymptotic distribution of the proposed test statistic is derived. The test is shown to be consistent against certain alternatives in the class of Gaussian processes.

2. THE TEST

Let $[x_1, \dots, x_N]$ be observations on a discrete linear stochastic process $[x_t]$ which is of the form

$$\begin{aligned} x_t &= y_t + m_t \\ y_t &= \sum_{\nu=-\infty}^{+\infty} a_{t-\nu} \xi_\nu \end{aligned} \quad \dots \quad (2.1)$$

$$E|\xi_\nu|^4 < \infty, \quad \sum_{-\infty}^{+\infty} |a_t|^2 < \infty$$

where ξ_ν are independent and identically distributed random variables, each with mean zero and variance one. Let

$$\Sigma = ((\rho_{i-j})) \quad \begin{aligned} 1 \leq i \leq N \\ 1 \leq j \leq N \end{aligned}$$

be the dispersion matrix of $[x_1, \dots, x_N]$. The function

$$f(\lambda) = \frac{1}{2\pi} \left| \sum_{-\infty}^{+\infty} a_t e^{it\lambda} \right|^2 \quad \dots \quad (2.2)$$

is the spectral density of the process. The problem is to test the null hypothesis

$$H_0 : m_t = 0 (t = \dots -1, 0, 1, \dots). \quad \dots \quad (2.3)$$

If the process is Gaussian the likelihood ratio test will lead to the statistic $x\Sigma^{-1}x'$, where $x = [x_1 x_2 \dots, x_N]$ and large values of the statistic are significant. The test based on this can be shown to be consistent against all alternatives for which

$$\lim_{N \rightarrow \infty} \frac{[m_1 \dots m_N] \Sigma^{-1} [m_1 \dots m_N]}{\sqrt{N}} = \infty \quad \dots (2.4)$$

But this statistic, even though its exact distribution is known, is not very suitable, since for large N it is difficult to compute Σ^{-1} . We suggest here a more convenient alternative procedure which also is consistent against alternatives of the type just considered.

Consider the expression

$$X = \frac{\sum_{t=1}^N b_t x_t}{\sqrt{\sum_s \sum_r b_r b_s \rho_{r-s}}} \quad \dots (2.5)$$

where $b_1, b_2, \dots, b_N$ are arbitrary constants. If all the x_t 's are of expectation zero we expect $|X_b|$ to be small. If we maximise $|X_b|^2$ over all $b_1, \dots, b_N$ we get $x \Sigma^{-1} x'$. Instead we rewrite (2.5) as

$$|X_b|^2 = \frac{\left| \frac{1}{2\pi} \int_{-\pi}^{\pi} h(\lambda) \frac{\sum_{r=1}^N x_r e^{-ir\lambda}}{f(\lambda)} f(\lambda) d\lambda \right|^2}{\int_{-\pi}^{\pi} |h(\lambda)|^2 f(\lambda) d\lambda} \quad \dots (2.6)$$

where

$$h(\lambda) = \sum_{r=1}^N b_r e^{ir\lambda}. \quad \dots (2.7)$$

Now we maximise (2.6) over all functions $h(\lambda)$, square integrable with respect to $f(\lambda)$, instead of over $h(\lambda)$ of the type (2.7). Assuming that $1/f(\lambda)$ is integrable over $(-\pi, \pi)$, we obtain thus the statistic

$$Y_N = \max_{h(\lambda)} |X_b|^2 = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \frac{\left| \sum_{r=1}^N x_r e^{ir\lambda} \right|^2}{f(\lambda)} d\lambda \quad \dots (2.8)$$

by an application of Schwartz's inequality.

The asymptotic null distribution of Y_N is given by the following theorem.

Theorem 1: Suppose that $a_t = O(t^{-\beta})$, $\beta > 3/2$ and $\int_{-\pi}^{\pi} [f(\lambda)]^{-1} e^{ir\lambda} d\lambda = O(r^{-1})$. Then, under the null hypothesis, the distribution of $\sqrt{N}[(Y_{N/N}) - 1]$ converges to the normal distribution with mean zero and variance σ^2 where $\sigma^2 = E\xi_t^2 - 1$.

ON TESTING THE MEAN OF A DISCRETE LINEAR PROCESS

Proof: Let

$$I_N = (1/2\pi N) \left| \sum_1^N x_t e^{it\lambda} \right|^2 \quad \dots (2.9)$$

$$I_{N,\xi} = (1/2\pi N) \left| \sum_1^N \xi_t e^{it\lambda} \right|^2$$

Then, by (2.1), if $m_t = 0$ for all t , we have

$$2\pi\sqrt{N} \int_{-\pi}^{\pi} [I_N(\lambda) - 2\pi f(\lambda) I_{N,\xi}(\lambda)] [f(\lambda)]^{-1} d\lambda = N^{-\frac{1}{2}} \sum_{s=-\infty}^{+\infty} \sum_{r=-\infty}^{+\infty} a_r a_s d_{rs} \dots (2.10)$$

where
$$d_{rs} = \sum_{m=1}^N \sum_{n=1}^N \xi_{n-r} \xi_{m-s} \rho^{(n-m)} - \sum_{m=1+s}^{N+s} \sum_{n=1+r}^{N+r} \xi_{n-r} \xi_{m-s} \rho^{(n-m)} \dots (2.11)$$

and
$$\rho^{(m-n)} = \int_{-\pi}^{+\pi} e^{i(n-m)\lambda} [f(\lambda)]^{-1} d\lambda. \dots (2.12)$$

There may be a lattice rectangle of points $R_{rs}^{(N)}$ of points (n, m) common to the sums involved in (2.11). The expressions corresponding to those points will cancel. Let $C_{rs}^{(N)}$ be the complement of $R_{rs}^{(N)}$ with respect to the set consisting of all the lattice points in both the summations. We then have

$$E|d_{rs}| \leq \sum_{(n,m) \in C_{rs}^{(N)}} [g(n-m)] \dots (2.13)$$

where
$$g(k) = |\rho^{(k)}|. \dots (2.14)$$

Since $\rho^{(k)} = O(k^{-1})$ and $a_t = O(t^{-\beta})$, $\beta > 3/2$, by the analysis carried out by Grenander and Rosenblatt (1951, chapter 6, pp. 191-192) we have

$$\sqrt{N} \int_{-\pi}^{\pi} [I_N(\lambda) - 2\pi f(\lambda) I_{N,\xi}(\lambda)] [f(\lambda)]^{-1} d\lambda \rightarrow 0$$

in probability as $N \rightarrow \infty$. We have

$$\sqrt{N} \left[\int_{-\pi}^{+\pi} I_{N,\xi}(\lambda) d\lambda - 1 \right] = \sqrt{N} \left[N^{-1} \sum_{t=1}^N \xi_t^2 - 1 \right]. \dots (2.15)$$

An application of the central limit theorem shows that the limit distribution of (2.15) is normal with mean zero and variance

$$\sigma^2 = E\xi_t^4 - 1.$$

This completes the proof.

3. CONSISTENCY OF THE TEST WHEN THE ξ_v ARE NORMALLY DISTRIBUTED

The proposed test of the null hypothesis H_0 is the following: Reject H_0 if and only if

$$\sigma^{-1}\sqrt{N}[(Y_{N/N})-1] > K_\alpha \quad \dots (3.1)$$

where K_α is the upper α percent point of the normal distribution with mean zero and variance unity.

Theorem 2: If ξ_v are normally distributed the test given by the critical region

$$(N/2)^{\frac{1}{2}} [(Y_{N/N})-1] > K_\alpha$$

where K_α is the α percent point of the normal distribution with mean zero and variance unity, is consistent against all alternatives $\{m_t\}$ ($t = \dots -1, 0, 1, 2, \dots$) which satisfy the condition (2.4).

Proof: Since $x\Sigma^{-1}x' \leq Y_N$, we have

$$\begin{aligned} P[\text{rejecting } H_0 | \{m_t\}] &= P[\sqrt{N/2}[(Y_{N/N})-1] > K_\alpha | \{m_t\}] \\ &\geq P\left[\sqrt{N/2}\left[\left(\frac{x\Sigma^{-1}x'}{N}\right)-1\right] > K_\alpha | \{m_t\}\right]. \quad \dots (3.3) \end{aligned}$$

But

$$x\Sigma^{-1}x' = (x-m)\Sigma^{-1}(x-m)' + 2m\Sigma^{-1}(x-m)' + m\Sigma^{-1}m'$$

where

$$m = [m_1, m_2, \dots, m_N].$$

Since the limiting distributions of

$$\sqrt{N/2} \left[\frac{(x-m)\Sigma^{-1}(x-m)'}{N} - 1 \right]$$

and

$$\frac{m\Sigma^{-1}(x-m)'}{(m\Sigma^{-1}m')^{\frac{1}{2}}}$$

exist as $N \rightarrow \infty$, and

$$\lim_{N \rightarrow \infty} \frac{m\Sigma^{-1}m'}{\sqrt{N}} = \infty,$$

it is easy to see that, in (3.3), the left side of the inequality within square brackets tends to infinity in probability. Thus the limit of (3.3) as $N \rightarrow \infty$ is unity. This completes the proof of Theorem 2.

REFERENCE

GRENNANDER, U. and ROSENBLATT, M. (1951): *Statistical Analysis of Stationary Time Series*. John Wiley & Sons, New York.

Paper received: January, 1960.

ON A PROBLEM OF BARTHOLOMEW IN LIFE TESTING

By P. S. SWAMY and S. A. D. C. DOSS

University of Poona, India

SUMMARY. Bartholomew's problem in life testing is considered under a more general model of Mendenhall and Hader viz., that the parent failure population is made up of two different sub-populations mixed in unknown proportions, each sub-population representing a different cause of failure. Maximum likelihood estimating equations of the parameters and the variance-covariance matrix of the estimates are obtained. Earlier well-known results in life testing are realised as particular cases of the present model.

1. INTRODUCTION

Until recently, the literature on life testing was concerned with estimation of parameters and testing procedures associated with observations from a single exponential distribution. Bartholomew (1957) considered, under the assumption, that all items that are put on life test come from the same exponential life distribution, the problem of obtaining the maximum likelihood estimate of the mean life when records are available giving the dates of installation on life test of n items of equipment, and the life test is terminated at some time T without reference to the experiment. Recently, Mendenhall and Hader (1958) considered the problem of estimation of parameters of mixed exponentially distributed failure populations. In the light of Mendenhall and Hader's general model Bartholomew's assumption looks often unrealistic. In practice, as was shown by Mendenhall and Hader (1958), there are often different causes of failure of an item and there are different distributions of failure representing the different causes of failure; and hence it is reasonable to assume that the parent distribution of failure of items is made up of two or more sub-populations of failure. The purpose of this note is to study Bartholomew's problem under the general model of mixed distributions of Mendenhall and Hader.

Many failure data, in practice, are seen to follow an exponential failure distribution (Davis, 1952). In the present note, following Mendenhall and Hader (1958), we assume the parent failure distribution to be made up of two sub-populations, each distributed as

$$f_i(x) = \frac{1}{\theta_i} e^{-x/\theta_i}, \quad x > 0, \quad \theta_i > 0, \quad i = 1, 2.$$

Let the two sub-populations be mixed in proportion $p : q$ ($p + q = 1$, $p, q \geq 0$). Then the frequency function of the parent population is given by

$$f(x) = \frac{p}{\theta_1} e^{-x/\theta_1} + \frac{q}{\theta_2} e^{-x/\theta_2}.$$

We assume that as soon as an item fails, the cause of failure is known and the sub-population from which the item comes is identified.

Without interfering unduly with the normal running of a factory process n items are put on life test at different points of time and the test is terminated at some time T . Let T_j be the time passed since the j -th item was put to test and t_{ij} be the length of life of that item. If $t_{ij} < T_j$ the j -th item is said to have failed, otherwise it is said to have survived. The probability of survival of j -th item is given by

$$\begin{aligned} Q_j &= p e^{-T_j/\theta_1} + q e^{-T_j/\theta_2}. \\ \text{Let } P_j &= 1 - Q_j. \end{aligned}$$

2. MAXIMUM LIKELIHOOD ESTIMATES OF THE PARAMETERS

If at the time of termination of the test r_i items from the i -th sub-population have failed, the likelihood of the sample is given by

$$L = C p^{r_1} q^{r_2} \prod_{j=1}^n Q_j^{1-a_{ij}} (f_1(t_{1j}))^{a_{1j}} (f_2(t_{2j}))^{a_{2j}} \quad \dots (2.1)$$

where $a_{ij} = 1$ or 0 according as the j -th item has failed or survived and C is a constant. The maximum likelihood estimates of θ_1 , θ_2 , and p are obtained by equating the partial derivatives of the logarithm of the likelihood, L , to zero and solving the equations simultaneously.

The maximum likelihood estimating equations are

$$\frac{\partial \log L}{\partial \theta_1} = \sum_{j=1}^n \left\{ k_j(1-a_{ij}) \frac{T_j}{\theta_1^2} + a_{1j} \left(\frac{t_{1j}}{\theta_1^2} - \frac{1}{\theta_1} \right) \right\} = 0, \quad \dots (2.2)$$

$$\frac{\partial \log L}{\partial \theta_2} = \sum_{j=1}^n \left\{ (1-k_j)(1-a_{ij}) \frac{T_j}{\theta_2^2} + a_{2j} \left(\frac{t_{2j}}{\theta_2^2} - \frac{1}{\theta_2} \right) \right\} = 0, \quad \dots (2.3)$$

$$\frac{\partial \log L}{\partial p} = \frac{r_1}{p} - \frac{r_2}{q} + \sum_{j=1}^n \frac{(k_j - p)}{pq} (1-a_{ij}) = 0$$

$$\text{where } \frac{\partial \log L}{\partial p} = \frac{p e^{-T_j/\theta_1}}{p e^{-T_j/\theta_1} + q e^{-T_j/\theta_2}} = 0. \quad \dots (2.4)$$

$Q_{ij} = e^{-T_j/\theta_i}$. The estimating equations can also be written as

$$\hat{p} = \frac{r_1 + \sum_{j=1}^n \hat{k}_j(1-a_{ij})}{n} \quad \dots (2.5)$$

$$\hat{\theta}_1 = \frac{1}{r_1} \sum_{j=1}^n \{ \hat{k}_j(1-a_{ij})T_j + a_{1j}t_{1j} \}, \quad \dots (2.6)$$

$$\hat{\theta}_2 = \frac{1}{r_2} \sum_{j=1}^n \{ (1-\hat{k}_j)(1-a_{ij})T_j + a_{2j}t_{2j} \}, \quad \dots (2.7)$$

$$\text{where, now, } \hat{k}_j = \frac{\hat{p}\hat{Q}_{1j}}{\hat{Q}_j} \quad \dots (2.8)$$

ON A PROBLEM OF BARTHOLOMEW IN LIFE TESTING

The simultaneous equations (2.5), (2.6) and (2.7) can be solved for the maximum likelihood estimates by using a suitably modified iterative process of Mendenhall and Hader (1958). By substituting $\hat{p}$, $\hat{\theta}_1$, $\hat{\theta}_2$ of equations (2.5), (2.6) and (2.7) into (2.8) it can be easily seen that equation (2.8) is of the form

$$\hat{k}_j = g_j(\hat{k}_1, \hat{k}_2, \dots, \hat{k}_n).$$

If a good first approximation to the vector $[\hat{k}_{10}, \hat{k}_{20}, \dots, \hat{k}_{n0}]$ is available, the modified iterative process is as follows. The accurate value of k_j is the solution of the equation $g_j - k_j = 0$. But when we are proceeding with an approximate $[\hat{k}_j]$ we have

$$g_j - k_j = D_j.$$

Let
$$g_j = \frac{1}{1+v_j} \quad \text{where } v_j = \frac{\hat{q}}{\hat{p}} \exp\left(\frac{T_j}{\hat{\theta}_1} - \frac{T_j}{\hat{\theta}_2}\right).$$

$$\frac{\partial D_j}{\partial \hat{k}_i} = \begin{cases} -g_j^2 \frac{\partial v_j}{\partial \hat{k}_i}, & 1 \neq j \\ -g_j^2 \frac{\partial v_j}{\partial \hat{k}_i} - 1, & 1 = j \end{cases} \quad \dots \quad (2.9)$$

where
$$\frac{\partial v_j}{\partial \hat{k}_i} = -v_j(1-a_{ij}) \left\{ \frac{1}{\hat{p}\hat{q}} + T_i T_j \left(\frac{1}{\hat{\theta}_1 r_1} + \frac{1}{\hat{\theta}_2 r_2} \right) \right\}.$$

The equation (2.9) can also be written as

$$-\frac{\partial D_j}{\partial \hat{k}_i} = g_j^2 \frac{\partial v_j}{\partial \hat{k}_i} + \delta_{ji},$$

where δ_{ij} is the Kronecker delta. Since

$$dD_j = - \sum_{i=1}^n \left(g_j^2 \frac{\partial v_j}{\partial \hat{k}_i} + \delta_{ji} \right) d\hat{k}_i,$$

now we have the relation

$$\begin{bmatrix} dD_1 \\ dD_2 \\ \vdots \\ dD_n \end{bmatrix} = - \left[g_j^2 \frac{\partial v_j}{\partial \hat{k}_i} + \delta_{ji} \right] \cdot \begin{bmatrix} d\hat{k}_1 \\ d\hat{k}_2 \\ \vdots \\ d\hat{k}_n \end{bmatrix}.$$

Assuming the matrix $\left[g_j^2 \frac{\partial v_j}{\partial \hat{k}_i} + \delta_{ji} \right]$ to be non-singular

we obtain

$$\begin{bmatrix} d\hat{k}_1 \\ d\hat{k}_2 \\ \vdots \\ d\hat{k}_n \end{bmatrix} = - \left[g_j^2 \frac{\partial v_j}{\partial \hat{k}_i} + \delta_{ji} \right]^{-1} \begin{bmatrix} dD_1 \\ dD_2 \\ \vdots \\ dD_n \end{bmatrix}$$

choosing $[dD_1, dD_2, \dots, dD_n] = -[D_{10}, D_{20}, \dots, D_{n0}]$

we have

$$\begin{bmatrix} \hat{k}_{11} \\ \hat{k}_{21} \\ \vdots \\ \hat{k}_{n1} \end{bmatrix} = \begin{bmatrix} \hat{k}_{10} \\ \hat{k}_{20} \\ \vdots \\ \hat{k}_{n0} \end{bmatrix} + \begin{bmatrix} d\hat{k}_{10} \\ d\hat{k}_{20} \\ \vdots \\ d\hat{k}_{n0} \end{bmatrix} = \begin{bmatrix} \hat{k}_{10} \\ \hat{k}_{20} \\ \vdots \\ \hat{k}_{n0} \end{bmatrix} + \left[g_j^2 \frac{\partial v_j}{\partial \hat{k}_i} + \delta_{ji} \right]^{-1} \cdot \begin{bmatrix} D_{10} \\ D_{20} \\ \vdots \\ D_{n0} \end{bmatrix} \dots (2.10)$$

Thus we get a second approximation to $[\hat{k}_j]$ viz., $[\hat{k}_{j1}]$. Putting this value of $[\hat{k}_{j1}]$ in equations (2.5), (2.6) and (2.7) we get the first approximate values of $\hat{p}$, $\hat{\theta}_1$, and $\hat{\theta}_2$. The iterative process can be repeated until the desired degree of accuracy is attained. $\hat{k}_{j0} = \frac{1}{2}$ for all j can serve well as a first approximation to $\hat{k}_j$. No definite inference can be drawn about the parameters when $r_i = 0$. However, it is easily seen that

$$\hat{\theta}_1 > \sum_{j=1}^n \hat{k}_j T_j, \quad \text{if } r_1 = 0$$

and

$$\hat{\theta}_2 > \sum_{j=1}^n (1 - \hat{k}_j) T_j, \quad \text{if } r_2 = 0.$$

3. PARTICULAR CASES OF THE PRESENT MODEL

Some well-known earlier results in life testing can, now, be obtained as particular cases of our model. In the present model $\hat{\theta}$ is the solution of the equation

$$\sum_{j=1}^n \left\{ k_j(1-a_{ij}) \frac{T_j}{\theta^2} + a_{ij} \left(\frac{t_{ij}}{\theta^2} - \frac{1}{\theta} \right) \right\} = 0.$$

We shall obtain, in the following, four special cases of our model :

(I) In the present model when $\theta_1 = \theta_2 = \theta$, $k_j = 1$ for all j and $p = 1$;

$$\hat{\theta} = \frac{1}{\sum_{j=1}^n a_j} \cdot \sum_{j=1}^n \{a_j t_j + (1-a_j) T_j\} = \frac{1}{r} \sum_{j=1}^n \{a_j t_j + (1-a_j) T_j\}.$$

This is the result of Bartholomew's (1957) model.

(II) When $\theta_1 = \theta_2 = \theta$ and $T_j = T$ for all the items (that is when all the n items are put to test simultaneously)

$$\hat{\theta} = \frac{1}{\sum_{j=1}^n a_j} \cdot \sum_{j=1}^n \{a_j t_j + (1-a_j) T\} = \frac{1}{r} \left\{ \sum_{j=1}^n t_j + (n-r) T \right\}.$$

ON A PROBLEM OF BARTHOLOMEW IN LIFE TESTING

This is a particular case of Bartholomew's model.

(III) When $\theta_1 = \theta_2 = \theta$ and there is no limit imposed on the truncation time but all the n items are put to test simultaneously and life test is terminated as soon as r items fail

$$\hat{\theta} = \frac{1}{r} \left\{ \sum_{j=1}^n t_j + (n-r)t_r \right\}.$$

This is the well-known result of Epstein and Sobel (1953). Their other results of (1954) can also be obtained as special cases of our model.

(IV) In the special case when $T_j = T$ for all j our model is essentially same as that of Mendenhall and Hader (1958).

4. VARIANCE-COVARIANCE MATRIX OF THE ESTIMATES

In the case of several parameters Fisher's information matrix is given by $[I_{\alpha\beta}]$ where $I_{\alpha\beta} = -E \left\{ \frac{\partial^2 \log L}{\partial \theta_\alpha \partial \theta_\beta} \right\}$. The variance-covariance matrix $[V_{\alpha\beta}]$ is given by the inverse matrix $[I^{\alpha\beta}]$.

The information matrix of $\hat{p}$, $\hat{\theta}_1$, and $\hat{\theta}_2$ is found to be

$$[I_{\alpha\beta}] = \begin{bmatrix} \sum_{j=1}^n \frac{pP_1(T_j)}{\theta_1^2} & \sum_{j=1}^n \frac{pqC_jT_j^2}{\theta_1^2\theta_2^2} & \sum_{j=1}^n \frac{-C_jT_j}{\theta_1^2} \\ & \sum_{j=1}^n \frac{qP_2(T_j)}{\theta_2^2} & \sum_{j=1}^n \frac{C_jT_j}{\theta_2^2} \\ & & \frac{1}{pq} \sum_{j=1}^n (1-C_j) \end{bmatrix}$$

where $\alpha, \beta = p, \theta_1, \text{ or } \theta_2$; $P_i(T_j) = 1 - e^{-T_j/\theta_i}$ and $C_j = \frac{k_j(1-k_j)Q_j}{pq}$.

Hence the variance-covariance matrix of the estimates $\hat{p}$, $\hat{\theta}_1$, and $\hat{\theta}_2$ is given by

$$[V_{\alpha\beta}] = \frac{1}{S} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix},$$

where

$$b_{11} = \frac{\theta_1^4}{p} \sum_{j,l=1}^n \{P_2(T_j)(1-C_l)\theta_2^2 - pC_jC_lT_jT_l\},$$

$$b_{22} = \frac{\theta_2^4}{q} \sum_{j,l=1}^n \{P_1(T_j)(1-C_l)\theta_1^2 - qC_jC_lT_jT_l\},$$

$$b_{33} = pq \sum_{j,l=1}^n \{P_1(T_j)P_2(T_l)\theta_1^2\theta_2^2 - pqC_jC_lT_j^2T_l^2\},$$

$$b_{12} = b_{21} = \theta_1^2\theta_2^2 \sum_{j,l=1}^n C_jT_j\{C_l(T_j-T_l)-T_j\},$$

$$b_{13} = b_{31} = q\theta_1^2 \sum_{j,l=1}^n C_jT_j\{pC_lT_l^2 + P_2(T_l)\theta_2^2\},$$

$$b_{23} = b_{32} = -p\theta_2^2 \sum_{j,l=1}^n C_jT_j\{qC_lT_l^2 + P_1(T_l)\theta_1^2\},$$

and

$$S = \sum_{j,l,m=1}^n C_jC_lT_jT_l[pq\{C_m(T_jT_l-T_lT_m-T_m^2)-T_jT_l\} - \\ - \{p\theta_1^2P_1(T_m) + q\theta_2^2P_2(T_m)\}] + \sum_{j,l,m=1}^n \theta_1^2\theta_2^2P_1(T_m)P_2(T_j)(1-C_l).$$

The results of the present note can be extended to a more general model where the number of sub-populations, i , is greater than two.

The authors have also obtained the estimates of the parameters and their variance-covariance matrix in the case when the parent failure population is made up of two sub-populations, one exponential and another normal.

ACKNOWLEDGEMENTS

The authors acknowledge their indebtedness to Dr. V. S. Huzurbazar, for his keen interest and constant guidance in carrying through the present work; to Professor K. V. Krishna Sastry and the referee for their helpful suggestions and comments, which have improved both the form and content of this note.

REFERENCES

- BARTHOLOMEW, D. J. (1957): A problem in life testing. *J. Amer. Stat. Ass.*, **52**, 350-355.
 DAVIS, D. J. (1952): An analysis of some failure data. *J. Amer. Stat. Ass.*, **47** 113-150.
 EPSTEIN, B. and SOBEL, M. (1953): Life testing. *J. Amer. Stat. Ass.*, **48**, 486-502.
 ——— (1954): Some theorems relevant to life testing from an exponential distribution. *Ann. Math. Stat.*, **25**, 373-381.
 MENDENHALL, W. and HADER, R. J. (1958): Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika*, **45**, 504-520.

Paper received: September, 1960.

STUDENTISATION OF TWO-STAGE SAMPLE MEANS FROM NORMAL POPULATIONS WITH UNKNOWN COMMON VARIANCE¹

By HAROLD RUBEN

*Columbia University*²

SUMMARY. The present paper derives a strengthened version of Stein's classic results concerning the joint estimation, with the aid of two-stage samples, of a set of means of normal populations with unknown common variance by confidence sets of predetermined dimensions and confidence coefficient, and, concomitantly, the testing of the means with a power function independent of the common variance.

The method used has certain advantages of directness and simplicity, and serves thereby to throw fresh light on the nature of the statistical inference employed. (This is discussed in some detail in the concluding section of the paper). A further advantage is that the method generalizes in a straightforward manner to deal with the analogous but more difficult problem of sampling from normal populations for which the variances are completely unknown. It is proposed to discuss the generalization in subsequent papers.

1. HISTORICAL NOTE

The general problem of obtaining sequential confidence sets was first posed by Wald (1947). Previously, however, Barnard (1946) had speculated that the extra degree of freedom conferred by sequential sampling might allow of the possibility of controlling the width between the lower and upper bounds of an estimate in interval estimation. At about the same time, Stein (1945) showed how a two-stage sampling procedure could be devised to yield confidence intervals of fixed length and confidence coefficient for the mean of a normal population of unknown variance and, equivalently to yield a test for the mean with power independent of the variance. By way of contrast, it had been shown some years earlier by Dantzig (1940) that these aims could not be achieved by means of samples of fixed size. More precisely, Dantzig demonstrated that any test on the mean based on a fixed number of observations, whose power function is to be independent of the variance must have constant power for all values of the mean equal to the size of the critical region. This result was extended to the case of the testing of the general linear hypothesis by Stein (1945) who, in the same paper, generalized his results on a single mean (referred to earlier) to include a two-stage sampling procedure which yields confidence sets of fixed volume and confidence coefficient for the set of means from a finite set of normal populations with unknown common variance and, equivalently, yields a test for the means with power independent of the common variance.

¹ This research was sponsored by the Office of Naval Research under Contract Number Nonr-266 (33), Project Number NR 042-034. Reproduction in whole or in part is permitted for any purpose of the United States Government.

² Now with The University, Sheffield, England.

In 1949, the present author (Ruben, 1950), working in ignorance of Stein's results rediscovered them in slightly stronger form by a different mode of argument which may have some advantages in the direction of greater simplicity, directness and flexibility, and may, therefore, serve to throw further light on the nature of the statistical inference employed. A further advantage is that the method generalizes quite naturally to deal with the more difficult problem of sampling from normal populations where the variances need not be equal and are unknown. Accordingly, confidence intervals or regions of predetermined dimensions and confidence coefficient for the means, or functions of the means, and similarly tests relating to the means with power independent of the unknown variances, are readily established (Ruben, 1950). In all these cases, the tests and estimation procedures are shown to be conservative which, however, leads to a certain loss of efficiency. The sampling procedure consists in obtaining preliminary samples which enable one to form estimates of the unknown variance or variances, these estimates determining a rule when to stop sampling at the second stage. From another point of view the loss in efficiency might then be said to result from the fact that the second-stage samples are not used to obtain information about the variances. This aspect is discussed in some detail in the concluding section of the paper.

The author's results relevant to the situation where the variances are assumed to be equal are reproduced here in order to make these more accessible to statisticians than has hitherto been the case, and are at the same time somewhat amplified in scope (though condensed in form). It is planned that this paper shall serve further as a convenient starting point for a subsequent paper which will deal with the case where the variances are unequal.

2. NOTATIONS

The following notations have been adopted in the text. Let Z be a random variable which admits of a non-zero probability density function in the range (z_1, z_2) . This function may involve a finite number of *known* constants, $a_1, a_2, \dots, a_p$. The quantity $z_{a_1, a_2, \dots, a_p; \alpha} \equiv z^*$ is defined by the equation

$$\alpha = \int_{z^*}^{z_2} \varphi_{a_1, a_2, \dots, a_p}(z) dz,$$

where $\varphi_{a_1, a_2, \dots, a_p}(z)$ is the density function of Z . Symbols such as $t_r; \alpha/2$, χ_r^2 , $F_{m, n; \alpha}$ etc. then have definite meanings.

Any random variable which is distributed as χ^2 on r degrees of freedom will generically be called a χ_r^2 . Similarly, we shall speak of a t_r , an $F_{m, n}$, etc.

TWO-STAGE SAMPLE MEANS FROM NORMAL POPULATIONS

Throughout the text, the distribution functions of a χ_r^2 and a t_r will be denoted by $F_r(\chi^2)$ and $G_r(t)$, respectively.

3. SPECIFICATION OF THE SAMPLING RULE AND THE EXPECTED SAMPLE SIZE

Supposing we are 'given' q unconnected normal populations with means μ_i ($i = 1, 2, \dots, q$) and unknown common variance σ^2 . We assume $\mu_i = \mu_i^0$ for $i = r+s+1, r+s+2, \dots, q$ ($s \geq 0, r+s \leq q$), and that the μ_i for $i = 1, 2, \dots, r+s$ are unknown. The two-stage sampling procedure to be discussed, denoted by $S(n_0, k)$, is characterised by two constants n_0 and k which must be chosen so as to meet the specifications imposed on the nature of the confidence sets or the power functions, but are otherwise arbitrary. Subject to these constraints then, the design constants n_0 and k should be chosen so as to maximise, as far as possible, the efficiency (in some sense) of the procedure in the class of all such two-stage sampling procedures. The maximisation of efficiency will not be considered here, but the reader is referred to Seelbinder (1953) who discussed the maximisation in one particular instance. In the sequel we shall be concerned with the estimation (in a confidence sense) and testing of the μ_i for $i = 1, 2, \dots, r$.

The sampling rule may be described as follows. Preliminary independent random samples, each of size n_0 , are drawn from the q populations, the result of such a sampling being a set of observed values x_{ij} ($i = 1, 2, \dots, q; j = 1, 2, \dots, n_0$). Further observations x_{ij} ($i = 1, 2, \dots, r; j = n_0+1, n_0+2, \dots, n$) are obtained from the first r populations, where n is determined by the formula

$$n = \max (\{k^2 s_0^2\}, n_0), \quad \dots \quad (3.1)$$

s_0^2 being an (unbiased) estimate of σ^2 on $\nu_0 = n_0 q - r - s$ degrees of freedom based on the preliminary observations,

$$s_0^2 = (n_0 q - r - s)^{-1} \left[\sum_{i=1}^{r+s} \sum_{j=1}^{n_0} x_{ij}^2 - \sum_{i=1}^{r+s} \left(\sum_{j=1}^{n_0} x_{ij} \right)^2 / n_0 + \sum_{i=r+s+1}^q \sum_{j=1}^{n_0} (x_{ij} - \mu_i^0)^2 \right] \quad \dots \quad (3.2)$$

and $\{c\}$ denotes the smallest integer not less than c . It will appear subsequently that $1/k$ is a studentising scale factor which plays the same role in this variable sample size procedure as the estimated standard errors of the means play in the more usual fixed sample size procedures.

The probability distribution of n is clearly a kind of lumped χ^2 distribution. In fact, from (3.1),

$$\begin{aligned} \Pr\{n = m | \mu, \sigma\} &= 0 & (m < n_0), \\ &= \Pr\{0 \leq s_0^2 \leq n_0/k^2\} & (m = n_0), \\ &= \Pr\{(m-1)/k^2 < s_0^2 \leq m/k^2\} & (m > n_0), \end{aligned} \quad \left. \vphantom{\Pr\{n = m | \mu, \sigma\}} \right\}$$

whence, denoting $\Pr\{n = m | \mu, \sigma\}$ by $p_m(\sigma)$,

$$\left. \begin{aligned} p_m(\sigma) &= 0 & (m < n_0) \\ &= F_{\nu_0}(n_0 h) & (m = n_0), \\ &= F_{\nu_0}(mh) - F_{\nu_0}((m-1)h) & (m > n_0), \end{aligned} \right\} \dots \quad (3.3)$$

where $h = \nu_0/k^2\sigma^2$.

Further, it may be readily shown from (3.3) (cf. Stein, 1945) that

$$\begin{aligned} n_0 F_{\nu_0}(n_0 h) + (\nu_0/h)[1 - F_{\nu_0+2}(n_0 h)] &< En < n_0 F_{\nu_0}(n_0 h) + (\nu_0/h)[1 - F_{\nu_0+2}(n_0 h)] + \\ &+ [1 - F_{\nu_0}(n_0 h)], \quad \dots \quad (3.4) \end{aligned}$$

$$\text{whence} \quad En \sim k^2\sigma^2, \quad \dots \quad (3.5)$$

the approximation being reasonably good when $\nu_0/h = k^2\sigma^2 > n_0$.

4. THE EXACT AND LIMITING PROBABILITY DISTRIBUTIONS OF THE TWO-STAGE SAMPLE MEANS AND OF FUNCTIONS OF THE MEANS

Denote the i -th two-stage sample mean by $\bar{x}_i$,

$$\bar{x}_i = \sum_{j=1}^n x_{ij}/n \quad (i = 1, 2, \dots, r),$$

and, for convenience, the vector $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_r)$ by $\bar{\mathbf{x}}$. The vector $(\mu_1, \mu_2, \dots, \mu_r)$ will similarly be denoted by $\boldsymbol{\mu}$. Then, for a given region R in r -space,

$$P(R | \boldsymbol{\mu}, \sigma) = \Pr\{\bar{\mathbf{x}} \in R | \boldsymbol{\mu}, \sigma\} = \sum_{m=1}^{\infty} p_m(\sigma) \Pr\{\bar{\mathbf{x}} \in R | n = m, \boldsymbol{\mu}, \sigma\}. \quad \dots \quad (4.1)$$

From the independence of s_0^2 and $\sum_{j=1}^{n_0} x_{ij}$ for $i = 1, 2, \dots, r$, it follows that the (conditional) distribution of $\bar{\mathbf{x}}$ for given n (equivalent, by (3.1), to a constraint on s_0^2) is multivariate normal with expectation vector $\boldsymbol{\mu}$ and variance-covariance matrix $(\sigma^2/m)\mathbf{I}$, where $\mathbf{I}$ is the unit $r \times r$ matrix. If then $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_r)$ is a multivariate normal vector with zero expectation vector and variance-covariance matrix $\mathbf{I}$, equation (4.1) may be stated as

$$P(R | \boldsymbol{\mu}, \sigma) = \sum_{m=1}^{\infty} p_m(\sigma) Q_m(R; \boldsymbol{\mu}, \sigma/\sqrt{m}), \quad \dots \quad (4.2)$$

where

$$Q_m(R; \boldsymbol{\mu}, \sigma/\sqrt{m}) = (2\pi)^{-\frac{1}{2}r} \int_{\boldsymbol{\mu} + (\sigma/\sqrt{m})\boldsymbol{\xi} \in R} \dots \int e^{-\frac{1}{2}\boldsymbol{\xi}'\boldsymbol{\xi}} d\boldsymbol{\xi}. \quad \dots \quad (4.3)$$

TWO-STAGE SAMPLE MEANS FROM NORMAL POPULATIONS

Equations (4.2) and (4.3) give the required *exact* probability distribution of $\bar{\mathbf{x}}$. In particular, these equations give the exact distribution function of any (Borel measurable) function, whether scalar or vector, of $\bar{\mathbf{x}}$. Specifically, if $y = \varphi(\bar{\mathbf{x}}) = \varphi(\bar{x}_1, \dots, \bar{x}_r)$ then the region R corresponding to

$$\Pr\{y_0 < \varphi(\bar{\mathbf{x}}) < y_1 | \boldsymbol{\mu}, \sigma\} \quad \dots \quad (4.4)$$

is defined by $y_0 < \varphi(\bar{\mathbf{x}}) < y_1$, and

$$Q_m(R; \boldsymbol{\mu}, \sigma/\sqrt{m}) = (2\pi)^{-\frac{1}{2}r} \int_{y_0 < \varphi(\boldsymbol{\mu} + (\sigma/\sqrt{m})\boldsymbol{\xi}) < y_1} \dots \int e^{-\frac{1}{2}\boldsymbol{\xi}'\boldsymbol{\xi}} d\boldsymbol{\xi}. \quad \dots \quad (4.5)$$

Similarly, if

$$\mathbf{y} = \boldsymbol{\varphi}(\bar{\mathbf{x}}) \equiv (\varphi_1(\bar{\mathbf{x}}), \varphi_2(\bar{\mathbf{x}}), \dots, \varphi_r(\bar{\mathbf{x}})),$$

then for

$$\Pr\{\mathbf{y}_0 < \boldsymbol{\varphi}(\bar{\mathbf{x}}) < \mathbf{y}_1 | \boldsymbol{\mu}, \sigma\} \quad \dots \quad (4.6)$$

$$\text{we have} \quad Q_m(R; \boldsymbol{\mu}, \sigma/\sqrt{m}) = (2\pi)^{-\frac{1}{2}r} \int_{\mathbf{y}_0 < \boldsymbol{\varphi}(\boldsymbol{\mu} + (\sigma/\sqrt{m})\boldsymbol{\xi}) < \mathbf{y}_1} \dots \int e^{-\frac{1}{2}\boldsymbol{\xi}'\boldsymbol{\xi}} d\boldsymbol{\xi}. \quad \dots \quad (4.7)$$

To determine the *limiting* probability distribution of $\bar{\mathbf{x}}$, observe that an alternative form of (4.2) is

$$P(R | \boldsymbol{\mu}, \sigma) = \sum_{m=1}^{\infty} Q_m(R; \boldsymbol{\mu}, \sqrt{v_0}/k\sqrt{u_m}) \Delta_m F_{v_0}(u), \quad \dots \quad (4.8)$$

where $u_m = mh$ and (referring to (3.3))

$$\left. \begin{aligned} \Delta_m F_{v_0}(u) &= 0 & (m < n_0), \\ &= F_{v_0}(u_{n_0}) & (m = n_0), \\ &= F_{v_0}(u_m) - F_{v_0}(u_{m-1}) & (m > n_0). \end{aligned} \right\} \quad \dots \quad (4.9)$$

Hence,

$$\begin{aligned} P^*(R | \boldsymbol{\mu}) &\equiv \lim_{\sigma \rightarrow \infty} P(R | \boldsymbol{\mu}, \sigma) = \lim_{h \rightarrow 0} \sum_{m=1}^{\infty} Q_m(R; \boldsymbol{\mu}, \sqrt{v_0}/k\sqrt{u_m}) \Delta_m F_{v_0}(u) \\ &= \int_0^{\infty} Q^*(R; \boldsymbol{\mu}, \sqrt{v_0}/k\sqrt{u}) dF_{v_0}(u), \quad \dots \quad (4.10) \end{aligned}$$

$$\text{where, by (4.3), } Q^*(R; \boldsymbol{\mu}, \sqrt{v_0}/k\sqrt{u}) = (2\pi)^{-\frac{1}{2}r} \int_{\boldsymbol{\mu} + (\sqrt{v_0}/k\sqrt{u})\boldsymbol{\xi} \in R} \dots \int e^{-\frac{1}{2}\boldsymbol{\xi}'\boldsymbol{\xi}} d\boldsymbol{\xi}. \quad \dots \quad (4.11)$$

Equations (4.10) and (4.11) give the limiting probability distribution of $\bar{\mathbf{x}}$ as $\sigma \rightarrow \infty$.

We remark that according to equations (4.10) and (4.11) the $\bar{x}_i$ ($i = 1, 2, \dots, r$) are distributed in the limit ($\sigma \rightarrow \infty$) as $\mu_i + \sqrt{v_0} \xi_i / k \sqrt{u}$, where $\xi_1, \xi_2, \dots, \xi_r$ and u are independent random variables, the ξ_i being normal with zero means and unit variances, while u is a $\chi^2_{v_0}$. For the right-hand member of (4.11) gives the (conditional) probability that the vector $\mu + (\sqrt{v_0}/k \sqrt{u}) \xi$ falls in R for a fixed value of u , and the right-hand member of (4.10) therefore gives the (unconditional) probability of this event, whatever be the value of u . Equivalently,

$$k(\bar{x} - \mu) \sim \xi / \sqrt{u/v_0}, \quad \dots \quad (4.12)$$

in which the sign $\sim$ in (4.12) is to be interpreted as meaning that the random vectors $k(\bar{x} - \mu)$ and $\xi / u^{1/2} v_0^{-1/2}$ have identical distributions in the limit as $\sigma \rightarrow \infty$. Similarly, from (4.10) and (4.11),

$$\varphi(\bar{x}) \sim \varphi(\mu + \xi / k \sqrt{u/v_0}) \quad \dots \quad (4.13)$$

and, more generally,
$$\varphi(\bar{x}) \sim \varphi(\mu + \xi / k \sqrt{u/v_0}). \quad \dots \quad (4.14)$$

Since $P^*(R|\mu) \neq 0$, equations (4.10) and (4.11) enable one, using the present two-stage sampling procedure $S(n_0, k)$, to obtain confidence regions of predetermined dimensions for the means, or functions of the means, and likewise to obtain tests relating to the means of predetermined power if σ^2 is regarded as large, i.e. the confidence assertions, as well as any assertions on the risks of error (of both the first and second kinds) will be *approximately* valid when the right-hand member of (4.10) is used to evaluate the appropriate probabilities, provided only that σ^2 is large. We now proceed to show that under certain conditions these assertions are necessarily of a conservative character. In other words, confidence probabilities and tests, based on the procedure $S(n_0, k)$, which use the right-hand member of (4.10) for the evaluation of probabilities will be better than stated. At the same time, the question of 'how much better' can to a considerable extent be decided by comparing $P^*(R|\mu)$ with $P(R|\mu, \sigma)$, where in the latter function σ is assessed by incorporating the entire weight of evidence, both *a priori* and *a posteriori*, from *all* the observations (not merely those in the pilot samples).

To prove the conservativeness of the statistical inference, observe that the right-hand member of equation (4.8) represents the upper (lower) Darboux sum associated with the division of the positive u -axis by the points $u_{n_0}, u_{n_0+1}, \dots$, and the integral occurring on the right-hand side of (4.10), if Q^* is strictly increasing (decreasing) in u . On recalling further that $u_m = mh$ ($m = n_0, n_0+1, \dots$) with $h = v_0/k^2\sigma^2$, and regarding $P(R|\mu, \sigma)$ as a function of h , it follows that the effect of a decrease of h by the factor λ (i.e. an increase of σ^2 by the factor λ), where λ is a positive integer > 1 , is to superimpose new points of division on the

TWO-STAGE SAMPLE MEANS FROM NORMAL POPULATIONS

original set of such points, and consequently to depress (increase) the Darboux sum. Therefore,¹ denoting for brevity $P(R|\mu, \sigma)$ by $\omega(h)$,

$$\left. \begin{aligned} \omega(h/\lambda) < \omega(h), Q^*(R; \mu, \sqrt{v_0}/k\sqrt{u}) \text{ strictly increasing in } u \\ (\lambda = 2, 3, \dots), \\ \omega(h/\lambda) > \omega(h), Q^*(R; \mu, \sqrt{v_0}/k\sqrt{u}) \text{ strictly decreasing in } u \\ (\lambda = 2, 3, \dots). \end{aligned} \right\} \dots \quad (4.15)$$

As an immediate corollary of (4.10) and (4.15), we obtain

$$\left. \begin{aligned} \int_0^\infty Q^*(R; \mu, \sqrt{v_0}/k\sqrt{u}) dF_{v_0}(u) &= \lim_{h \rightarrow 0} \omega(h) = \inf_h \omega(h), \\ &Q^* \text{ strictly increasing in } u \\ &= \sup_h \omega(h), Q^* \text{ strictly decreasing in } u, \end{aligned} \right\} \dots \quad (4.16)$$

or, equivalently, the greatest lower bound (least upper bound) of $P(R|\mu, \sigma)$ is attained asymptotically as $\sigma \rightarrow \infty$ if $Q_m(R; \mu, \sigma/\sqrt{m})$ is strictly decreasing (increasing) in σ .

To verify the right half of (4.16) suppose that $Q^*(R; \mu, \sqrt{v_0}/k\sqrt{u})$ is strictly increasing in u (the argument for strictly decreasing Q^* is similar), and assume the existence of a value of h , $h=h'$ (say), for which $\omega(h') < \lim_{h \rightarrow 0} \omega(h)$. Let $\lim_{h \rightarrow 0} \omega(h) - \omega(h') = c (> 0)$. Then, by (4.15), $\lim_{h \rightarrow 0} \omega(h) - \omega(h'/\lambda) > c$, for $\lambda = 2, 3, \dots$, i.e. infinitely many values of h can be found in every arbitrarily small neighbourhood round $h = 0$ for which the corresponding values of $\omega(h)$ differ from their limiting value, $\lim_{h \rightarrow 0} \omega(h)$, by more than the given quantity c . But this contradicts the existence of the latter limit whose value has, however, been previously determined and is given in the left half of equation (4.16). Thus, no such value of h as postulated exists, and $\inf \omega(h)$ is given by the limiting value $\lim_{h \rightarrow 0} \omega(h)$. Equation (4.16) is, therefore, proved.

5. SOME EXAMPLES AND APPLICATIONS

(A) *Two-stage sampling from a single normal population.* Consider two-stage sampling from a single normal population ($q = 1 = r, s = 0$) with unknown mean μ and variance σ^2 , respectively, according to the scheme in (3.1),

i.e.
$$n = \max \{k^2 s_0^2, n_0\},$$

with
$$s_0^2 = (n_0 - 1)^{-1} \left[\sum_j^{n_0} x_j^2 - \left(\sum_1^{n_0} x_j \right)^2 / n_0 \right]. \quad \dots \quad (5.1)$$

Here,
$$v_0 = n_0 - 1.$$

¹It seems highly likely on intuitive grounds that $\omega(h)$ is strictly increasing (decreasing) in h if $Q^*(R; \mu, \sqrt{v_0}/k\sqrt{u})$ is strictly increasing (decreasing) in u . However, a formal proof of this has not been found in the general case.

From (4.12),

$$k(\bar{x}-\mu) \sim t_{n_0-1}, \quad \dots \quad (5.2)$$

where

$$\bar{x} = \sum_1^n x_j/n. \quad \dots \quad (5.3)$$

It follows from (5.2) that to obtain (say) upper confidence intervals for μ of the form $(\bar{x}-a, \infty)$ with confidence coefficient $1-\alpha$, where a and α are preassigned quantities (these intervals having the property that their lower end-points are to the left of μ with a probability of $1-\alpha$), the design constants n_0 and k must be related to the specification constants a and α by the equation

$$ak = t_{n_0-1}; \alpha. \quad \dots \quad (5.4)$$

The confidence intervals are based on the probability relation

$$P\{k(\bar{x}-\mu) < t_{n_0-1}; \alpha | \mu\} \sim 1-\alpha. \quad \dots \quad (5.5)$$

The single constraint (5.4) on the design constants fails to fix them uniquely but on the contrary provides one degree of freedom for their choice. A rational choice of n_0 and k subject to (5.4) may be obtained as follows. Use (3.3) and (5.4) to plot $En = \sum m p_m(\sigma)$ as a function of σ , for fixed a and α , and for varying n_0 , so generating a family of curves; if σ is suspected *a priori* to have an approximate value of σ' , choose that value of n_0 from the family of curves of En which will minimise En for $\sigma = \sigma'$. (Cf. Seelbinder, 1953).

Similarly, to test the one-sided null hypothesis $\mu = \mu_0$ on significance level α against the set of alternatives $\mu > \mu_0$ the critical region is, from (5.2), defined by

$$k(\bar{x}-\mu_0) \geq t_{n_0-1}; \alpha, \quad \dots \quad (5.6)$$

and, since $k(\bar{x}-\mu_0) = k(\bar{x}-\mu) + k(\mu-\mu_0) = t_{n_0-1} + k(\mu-\mu_0), \quad \dots \quad (5.7)$

the limiting power ($\sigma \rightarrow \infty$) of this test is

$$1 - G_{n_0-1} [t_{n_0-1}; \alpha - k(\mu-\mu_0)]. \quad \dots \quad (5.8)$$

The power function is naturally strictly increasing in $\mu-\mu_0$, and the test is insensitive to values of $\mu < \mu_0$.

For all real and positive values of k , once n_0 is chosen there are an infinity of rejection criteria, and a corresponding infinity of power functions given by (5.8). To fix the test in any instance, additional information other than that the size of the critical region is α is needed. In fact, when n_0 has been chosen, *two* points on the power curve are needed to fix the latter completely. A further imposed condition might be that when $\mu \geq \mu_0 + \delta$, the probability of an error of the second kind should not exceed

TWO-STAGE SAMPLE MEANS FROM NORMAL POPULATIONS

β , where δ is a given positive number and β is some given (small) positive quantity (say, 0.05 or 0.01). We then require that

$$\beta = G_{n_0-1}(t_{n_0-1}; \alpha - k\delta), \quad \dots \quad (5.9)$$

giving
$$t_{n_0-1}; \alpha - k\delta = -t_{n_0-1}; \beta, \quad \dots \quad (5.10)$$

whence
$$k = (t_{n_0-1}; \alpha + t_{n_0-1}; \beta) / \delta \quad \dots \quad (5.11)$$

and, by (3.5),
$$En \sim \left(\frac{t_{n_0-1}; \alpha + t_{n_0-1}; \beta}{\delta} \right)^2 \sigma^2. \quad \dots \quad (5.12)$$

As for the confidence interval case, n_0 and k should be chosen, subject to (5.10), so as to minimise En around the true suspected value of σ .

The estimation and test procedures are both conservative. Referring to (5.5) and (4.3) the value of Q_m for the estimation procedure is

$$\Phi \left(\frac{t_{n_0-1}; \alpha}{k} \frac{\sqrt{m}}{\sigma} \right), \quad \dots \quad (5.12)$$

where $\Phi(\cdot)$ denotes, as usual, the distribution function of a normal variate with zero mean and unit variance. Since Q_m is decreasing in σ , the value of the left-hand member of (5.5), and, therefore, also the confidence coefficient, is greater than $1 - \alpha$, whatever be the 'true' value of σ . The true value of the confidence coefficient is

$$\sum_1^{\infty} p_m(\sigma) \Phi \left(\frac{t_{n_0-1}; \alpha}{k} \frac{\sqrt{m}}{\sigma} \right). \quad \dots \quad (5.13)$$

Again, on referring to (5.6) and (4.3), we find that the value of Q_m for the test procedure is

$$1 - \Phi \left[\left(\frac{t_{n_0-1}; \alpha}{k} - (\mu - \mu_0) \right) \frac{\sqrt{m}}{\sigma} \right]. \quad \dots \quad (5.14)$$

This is strictly increasing or decreasing in σ according as to whether $0 \leq \mu - \mu_0 < t_{n_0-1}; \alpha/k$ or $\mu - \mu_0 > t_{n_0-1}; \alpha/k$, and is identically equal to $\frac{1}{2}$ when $\mu - \mu_0 = t_{n_0-1}; \alpha/k$. Hence, the true power of the test, given by

$$\sum_1^{\infty} p_m(\sigma) \left\{ 1 - \Phi \left[\left(\frac{t_{n_0-1}; \alpha}{k} - (\mu - \mu_0) \right) \frac{\sqrt{m}}{\sigma} \right] \right\}, \quad \dots \quad (5.15)$$

is likewise increasing in σ if $0 \leq \mu - \mu_0 < t_{n_0-1}; \alpha/k$ and decreasing in σ if $\mu - \mu_0 > t_{n_0-1}; \alpha/k$, while it is identically equal to $\frac{1}{2}$ when $\mu - \mu_0 = t_{n_0-1}; \alpha/k$ (i.e. the test is insensitive to variations in σ when μ lies in the neighbourhood of $\mu_0 + t_{n_0-1}; \alpha/k$). In brief, the true power of the test, as given by (5.15), is better than the limiting power in (5.8), obtained by letting $\sigma \rightarrow \infty$, in the sense that it is less than the limiting power for small or moderate μ and greater than the limiting power for large μ .

The procedures for obtaining two-sided confidence intervals and tests are quite similar. Thus, to obtain two-sided, symmetrical confidence intervals for μ of preassigned length $2a$ and confidence coefficient $1-\alpha$, the constants n_0 and k are related to a and α by

$$ak = t_{n_0-1; \alpha/2}, \quad \dots (5.16)$$

the confidence intervals being $(\bar{x}-a, \bar{x}+a)$, and based on the probability relation

$$\Pr \{ |k(\bar{x}-\mu)| < t_{n_0-1; \alpha/2} | \mu \} \sim 1-\alpha. \quad \dots (5.17)$$

In the same way, to test $\mu = \mu_0$ against $\mu \neq \mu_0$ on significance level α , the critical region is

$$k(\bar{x}-\mu_0) \geq t_{n_0-1; \alpha/2}. \quad \dots (5.18)$$

The limiting power of the test, by (5.2) and (5.18), is

$$1 - [G_{n_0-1}(t_{n_0-1; \alpha/2} - k(\mu - \mu_0)) - G_{n_0-1}(-t_{n_0-1; \alpha/2} - k(\mu - \mu_0))]. \quad \dots (5.19)$$

The test is clearly unbiased and its power function is monotonic increasing in $|\mu - \mu_0|$.

To fix the test and power function uniquely once n_0 has been chosen, an additional specification is required of the type that when $|\mu - \mu_0| \geq \delta$, the probability of an error of the second kind does not exceed β , δ being a fixed, given positive quantity and β a fixed, given (small) positive number (say, 0.05 or 0.01). Thus,

$$\beta = G_{n_0-1}(t_{n_0-1; \alpha/2} - k\delta) - G_{n_0-1}(-t_{n_0-1; \alpha/2} - k\delta), \quad \dots (5.20)$$

from which k may be determined numerically or graphically for given α, β, δ and for varying n_0 . (The right-hand member of (5.20) is strictly decreasing in k and ranges in value from $1-\alpha$ when $k=0$ to zero as $k \rightarrow \infty$.) In principle, n_0 and k should be chosen, subject to (5.20), so as to minimise En , as before, around the true value of σ .

The estimation and test procedures are again conservative. For the estimation procedure, Q_m is (refer to (5.17) and (4.3))

$$2\Phi\left(\frac{a\sqrt{m}}{\sigma}\right) - 1, \quad \dots (5.21)$$

which is decreasing in σ . Thus, the left-hand member of (5.17) is likewise decreasing in σ and the confidence coefficient is, therefore, greater than $1-\alpha$, irrespective of σ . The true confidence coefficient is

$$\sum_1^{\infty} p_m(\sigma) \left[2\Phi\left(\frac{a\sqrt{m}}{\sigma}\right) - 1 \right]. \quad \dots (5.22)$$

TWO-STAGE SAMPLE MEANS FROM NORMAL POPULATIONS

For the test procedure, Q_m is

$$\Phi \left[(t_{n_0-1}, \alpha/2/k - (\mu - \mu_0)) \frac{\sqrt{m}}{\sigma} \right] - \Phi \left[(-t_{n_0-1}, \alpha/2/k - (\mu - \mu_0)) \frac{\sqrt{m}}{\sigma} \right] \dots \quad (5.23)$$

This function increases steadily with σ for $0 \leq |\mu - \mu_0| < t_{n_0-1}, \alpha/2/k$ and decreases steadily with σ for sufficiently large σ , while for moderate values of $|\mu - \mu_0|$ it is hardly affected by variations in σ . Hence, the true power of the test, given by

$$\sum_1^{\infty} p_m(\sigma) \left\{ \Phi \left[(t_{n_0-1}, \alpha/2/k - (\mu - \mu_0)) \frac{\sqrt{m}}{\sigma} \right] - \Phi \left[(-t_{n_0-1}, \alpha/2/k - (\mu - \mu_0)) \frac{\sqrt{m}}{\sigma} \right] \right\}, \dots \quad (5.24)$$

is better than the limiting value given in (5.19).

(B) *Two-stage sampling from a finite set of normal populations with common variance.* Consider q separate normal populations with unknown common variance σ^2 , and with means given thus:

$$\left. \begin{aligned} Ex_i &= \mu_i, \quad i = 1, 2, \dots, r+s(r+s \leq q, s \geq 0), \\ &= 0, \quad i = r+s+1, r+s+2, \dots, q. \end{aligned} \right\} \dots \quad (5.25)$$

Hyperspherical confidence regions of predetermined volume content and confidence coefficient $1-\alpha$ may be obtained from the set of means μ_i ($i = 1, 2, \dots, r$) if sampling is carried out in two stages, as in Section 1, according to the scheme of equation (3.1) with

$$s_0^2 = (n_0q - r - s)^{-1} \left[\sum_{i=1}^q \sum_{j=1}^{n_0} x_{ij}^2 - \sum_{i=1}^{r+s} \left(\sum_{j=1}^{n_0} x_{ij} \right)^2 / n_0 \right], \dots \quad (5.26)$$

the degrees of freedom of s_0^2 being $\nu_0 = n_0q - r - s$.

On setting

$$\varphi(\bar{\mathbf{x}}) = (\bar{\mathbf{x}} - \boldsymbol{\mu})' (\bar{\mathbf{x}} - \boldsymbol{\mu})$$

in (4.13), with (as before) $\bar{\mathbf{x}} = (\bar{x}_1, \dots, \bar{x}_r)$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_r)$, we have

$$(\bar{\mathbf{x}} - \boldsymbol{\mu})' (\bar{\mathbf{x}} - \boldsymbol{\mu}) \sim \frac{1}{k^2} \frac{\boldsymbol{\xi}' \boldsymbol{\xi}}{u/\nu_0}, \dots \quad (5.27)$$

or

$$\sum_1^r (\bar{x}_i - \mu_i)^2 \sim \frac{1}{k^2} \frac{\sum_1^r \xi_i^2}{u/\nu_0}, \dots \quad (5.28)$$

where the components of $\boldsymbol{\xi}$ are independent normal random variables with zero means and unit variances, u is a $\chi_{\nu_0}^2$ and $\boldsymbol{\xi}$ and u are independently distributed. Thus,

$$\frac{k^2}{r} \sum_1^r (\bar{x}_i - \mu_i)^2 \sim F_{r, \nu_0}, \dots \quad (5.29)$$

and the required confidence regions are defined by

$$\frac{k^2}{r} \sum_1^r (\bar{x}_i - \mu_i)^2 < F_{r, v_0; \alpha}, \quad \dots \quad (5.30)$$

the corresponding probability relationship being

$$\Pr \left\{ \frac{k^2}{r} \sum_1^r (\bar{x}_i - \mu_i)^2 < F_{r, v_0; \alpha} \mid \mu \right\} \sim 1 - \alpha. \quad \dots \quad (5.31)$$

It follows from (5.29) that to obtain confidence regions for the $\mu_i (i = 1, 2, \dots, r)$ of the form

$$\sum_1^r (\bar{x}_i - \mu_i)^2 < \delta \quad \dots \quad (5.32)$$

with confidence coefficient $1 - \alpha$, δ and α being specified in advance (this region representing the interior of a random hypersphere in the space of the $\mu_i (i = 1, 2, \dots, r)$ with centre $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_r)$ and fixed radius $\sqrt{\delta}$), the following relation must hold:

$$k^2 \delta = r F_{r, v_0; \alpha}. \quad \dots \quad (5.33)$$

The fixed sample size n_F required to obtain a confidence region (interior of hypersphere) of the same volume and with the same confidence coefficient for *known* σ is given by

$$\Pr \{ n_F \sum_1^r (\bar{x}_i - \mu_i)^2 / \sigma^2 < \delta n_F / \sigma^2 \} = 1 - \alpha, \quad \dots \quad (5.34)$$

whence

$$\delta n_F / \sigma^2 = \chi_{r; \alpha}^2. \quad \dots \quad (5.35)$$

Provided then that the σ^2 is not excessively small, the efficiency of the two-stage procedure is, by (5.35), (5.33) and (3.5),

$$\frac{n_F}{En} = \frac{\chi_{r; \alpha}^2}{r F_{r, v_0; \alpha}} = \frac{F_{r, \infty; \alpha}}{F_{r, v_0; \alpha}}. \quad \dots \quad (5.36)$$

The same remarks about a rational choice of n_0 and k , subject to the constraint (5.33), as were made previously in relation to the simpler one-population case apply here with equal force, the purpose being once again to minimise the average amount of sampling, as far as possible.

Finally, in this instance (refer to (5.32), (5.33) and (4.3))

$$Q_m = F_{v_0} \left(\frac{r F_{r, v_0; \alpha}}{k^2} \frac{m}{\sigma^2} \right). \quad \dots \quad (5.37)$$

TWO-STAGE SAMPLE MEANS FROM NORMAL POPULATIONS

Since Q_m is decreasing in σ , so is the left-hand member of (5.31) and the true confidence coefficient actually exceeds $1-\alpha$. The true confidence coefficient is

$$\sum_1^{\infty} p_m(\sigma) F_{v_0} \left(\frac{r F_{r, v_0; \alpha} m}{k^2 \sigma^2} \right). \quad \dots (5.38)$$

We now proceed to discuss the testing for the set of means, specifically to test for the general linear hypothesis. We assume that the linear hypothesis has first been reduced to canonical form. The problem may then be formulated as follows. Given $q(\geq 1)$ separate normal populations with unknown common variance σ^2 and with means as in (5.25), we require to test the hypothesis $H_0: \mu_i = 0$ ($i = 1, 2, \dots, r$) with power to be (effectively) independent of σ^2 .

This can be achieved if sampling is carried out in two stages, exactly as in Section 3, i.e. equations (3.1) and (5.26) are still valid. The test procedure consists in rejecting H_0 when

$$\sum_1^r \bar{x}_i^2 \geq \frac{r}{k^2} F_{r, v_0; \alpha}, \quad \dots (5.39)$$

where $v_0 = n_0 q - r - s$, the corresponding probability relationship being

$$\Pr \left\{ \frac{k^2}{r} \sum_1^r \bar{x}_i^2 \geq F_{r, v_0; \alpha} \mid H_0 \right\} \sim \alpha. \quad \dots (5.40)$$

To determine the limiting power ($\sigma \rightarrow \infty$) of this test, we need the limiting distribution of $\bar{\mathbf{x}}' \mathbf{x}' = \sum_1^r \bar{x}_i^2$ for arbitrary $\boldsymbol{\mu} = (\mu_1, \dots, \mu_r)$. Let the region of Section 4 be defined by

$$R: \sum_1^r \bar{x}_i^2 < \delta, \quad \dots (5.41)$$

for arbitrary δ . Then,

$$\begin{aligned} Q_m(R; \boldsymbol{\mu}, \sigma/\sqrt{m}) &= (2\pi)^{-\frac{1}{2}r} \int \dots \int_{\boldsymbol{\mu} + (\sigma/\sqrt{m})\boldsymbol{\xi} \in R} e^{-\frac{1}{2}\boldsymbol{\xi}'\boldsymbol{\xi}} d\boldsymbol{\xi} \\ &= (2\pi)^{-\frac{1}{2}r} \int \dots \int_{\sum_1^r (\mu_i + \sigma \xi_i / \sqrt{m})^2 < \delta} e^{-\frac{1}{2}\boldsymbol{\xi}'\boldsymbol{\xi}} d\boldsymbol{\xi} \\ &= (2\pi)^{-\frac{1}{2}r} \int \dots \int_{\sum_1^r (\xi_i + \mu_i \sqrt{m}/\sigma)^2 < m\delta/\sigma^2} e^{-\frac{1}{2}\boldsymbol{\xi}'\boldsymbol{\xi}} d\boldsymbol{\xi} \\ &= \psi_r \left(\frac{m\delta}{\sigma^2}; \sqrt{\frac{m}{\sigma^2} \sum_1^r \mu_i^2} \right), \quad \dots (5.42) \end{aligned}$$

where $\psi_r(v; \kappa)$ is the distribution function of a non-central χ^2 with r degrees of freedom and with non-centrality parameter κ , i.e. the distribution of $\sum_1^r (\xi_i - c_i)^2$ with $+\sqrt{\sum_1^r c_i^2} = \kappa$. The limiting distribution function of $\sum_1^r \bar{x}_i^2$ is therefore, by (4.10),

$$\beta^*(\lambda; \rho) = \int_0^\infty \psi_r(\rho u; \sqrt{\lambda u}) dF_{v_0}(u), \quad \dots (5.43)$$

where $\rho = k^2 \delta / v_0$, $\lambda = k^2 \sum_1^r \mu_i^2 / v_0$. Since the density function of the non-central χ^2 is

$$\begin{aligned} \frac{\partial}{\partial v} \psi_r(v; \kappa) &= \frac{1}{2} \kappa^{-(\frac{1}{2}r-1)} e^{-\frac{1}{2}\kappa^2} e^{-\frac{1}{2}v} v^{\frac{1}{2}(\frac{1}{2}r-1)} I_{\frac{1}{2}r-1}(\kappa\sqrt{v}) \\ &= \frac{1}{2^{\frac{1}{2}r}} e^{-\frac{1}{2}\kappa^2} v^{\frac{1}{2}r-1} e^{-\frac{1}{2}v} \sum_{l=0}^\infty \frac{1}{\Gamma(l + \frac{r}{2})} \left(\frac{\kappa\sqrt{v}}{2} \right)^{2l} / l! \quad \dots (5.44) \end{aligned}$$

(see for instance, Patnaik, 1949), I denoting the Bessel Function of the first kind with purely imaginary argument, differentiation under the integral sign in (5.43) with respect to ρ yields (after a little algebra)

$$\frac{\partial}{\partial \rho} \beta^*(\lambda; \rho) = \sum_{l=0}^\infty \frac{\Gamma\left(\frac{n_0 q - r - s}{2} + 2l\right)}{\Gamma\left(\frac{r}{2} + l\right) \Gamma\left(\frac{n_0 q - r - s}{2}\right)} \frac{\lambda^l \rho^{\frac{1}{2}r+l-1}}{l!(1+\lambda+\rho)^{\frac{1}{2}(n_0 q - r - s) + 2l}} \cdot \dots (5.45)$$

On integrating term by term in (5.45) with respect to ρ , we find

$$\begin{aligned} \beta^*(\lambda; \rho) &= (1+\lambda)^{-\frac{1}{2}(n_0 q - r - s)} \sum_{l=0}^\infty \frac{(\lambda/1+\lambda)^l}{l!} \times \\ &\times \frac{\Gamma\left(\frac{n_0 q - r - s}{2} + l\right)}{\Gamma\left(\frac{n_0 q - r - s}{2}\right)} I_{\rho/(1+\lambda+\rho)}\left(\frac{r}{2} + l, \frac{n_0 q - r - s}{2} + l\right), \quad \dots (5.46) \end{aligned}$$

where in (5.46) I denotes the Incomplete Beta Function ratio. Note that

$$\beta^*(\lambda; \rho) = \beta^*(k^2 \sum_1^r \mu_i^2 / (n_0 q - r - s); r F_{n_0 q - r - s; \alpha} / (n_0 q - r - s)) \quad \dots (5.47)$$

also gives, by (5.39), the limiting probability of an error of the second kind,

$$\lim_{\sigma \rightarrow \infty} \Pr \left\{ \sum_1^r \bar{x}_i^2 < \delta \mid \mu, \sigma \right\}, \delta = \frac{r}{k^2} F_{r, n_0 q - r - s; \alpha}.$$

The limiting power function is, of course, $1 - \beta^*(\lambda; \rho)$.

TWO-STAGE SAMPLE MEANS FROM NORMAL POPULATIONS

It is easily seen that $\partial \psi_r(v; \kappa) / \partial \kappa < 0$. Hence, from (5.43), the limiting power function is strictly increasing in λ , and the test is unbiased. In practice, we should, as a rule, require only that

$$\beta^*(\lambda'; rF_{n_0q-r-s}; \alpha/(n_0q-r-s)) = \beta', \quad \dots \quad (5.48)$$

where λ' and β' are given positive quantities and β' will, as a rule, be small. The latter condition can be met by an appropriate unique choice of the scale constant $1/k$ for every fixed n_0 , and analogously to the previous examples the choice of n_0 and k should be such as to minimise En around the most likely value of σ .

We remark that Stein (1945) gives an integral form* for the power function which may be shown to be equivalent to the more convenient series expansion in (5.46).

The statements made in connexion with the limiting value of the power function for the two-stage test of Student's hypothesis may be, in essentials, duplicated here. The test is slightly better than is expressed by (5.47), in the sense that for all finite σ the true power function, which by (5.39) and (5.42) is given by

$$\begin{aligned} & 1 - \beta(\lambda; rF_{n_0q-r-s}; \alpha/(n_0q-r-s); \sigma) \\ &= 1 - \sum_1^{\infty} p_m(\sigma) \psi_r \left(\frac{F_{r, n_0q-r-s}; \alpha}{k^2} \frac{m}{\sigma^2}; \sqrt{\frac{m}{\sigma^2} \sum_1^r \mu_i^2} \right), \quad \dots \quad (5.49) \end{aligned}$$

satisfies the inequalities

$$\left. \begin{aligned} \beta(\lambda; rF_{n_0q-r-s}; \alpha/(n_0q-r-s); \sigma) &> \beta^*(\lambda; rF_{n_0q-r-s}; \alpha/(n_0q-r-s)), \quad 0 \leq \lambda \leq \lambda'', \\ &< \beta^*(\lambda; rF_{n_0q-r-s}; \alpha/(n_0q-r-s)), \quad \lambda > \lambda''', \end{aligned} \right\} \quad \dots \quad (5.50)$$

where $\lambda''' > \lambda''$, while for values of λ in the range (λ'', λ''') the power function is more or less insensitive to variations in σ . The inequalities in (5.50) derive from the monotonic increasing character of $\psi_r(\rho u; \sqrt{\lambda u})$, and therefore of Q^* , in u when λ is small. (Q^* is defined in (4.11)). Thus, for $\lambda = 0$, it has been established immediately after (5.37) that

$$\beta(0; rF_{r, n_0q-r-s}; \alpha/(n_0q-r-s); \sigma) > \beta^*(0; rF_{n_0q-r-s}; \alpha/(n_0q-r-s)),$$

and general considerations of continuity show that $\psi_r(\rho u; \sqrt{\lambda u})$ must be increasing in u near $u = 0$. On the other hand, $\psi_r(\rho u; \sqrt{\lambda u})$ is the total probability mass, under a spherical normal distribution with zero means and unit variances, in a

*Stein's integral is *not* the integral we have used, viz., $1 - \int_0^{\infty} \psi_r(\rho u; \sqrt{\lambda u}) du$, to obtain the series expansion (5.46) though, of course, the two integrals must be equivalent.

hypersphere with centre at $\sqrt{\lambda u}$ (which may for convenience be taken to lie along one of the coordinate axes) and radius $\sqrt{\rho u}$. Hence, it is clear that an infinitesimally small increase in u will affect the location of the centre (which moves away from the centre of the distribution) more strongly than the length of radius, when λ is sufficiently large. A net decrease of the probability mass accordingly results, i.e., $\partial \psi_r(\rho u; \sqrt{\lambda u}) / \partial u < 0$ for sufficiently large λ . This property may also be proved analytically (if rather tediously).

(C) *Joint one-sided confidence intervals for a finite set of means.* Given the populations as in 5(B), with sampling carried out in two stages also as in 5(B), confidence regions for the vector $\mu = (\mu_1, \mu_2, \dots, \mu_r)$ of the form $(\bar{x} - a, \infty)$ (i.e. infinitely extended orthotopes) with confidence coefficient $1 - \alpha$, where $a = (a_1, a_2, \dots, a_r)$ and α are predetermined may be obtained analogously to the one-sided confidence intervals for the case of a single population (though with greater computational difficulties because of the lack of tables for the distribution of the ensuing statistics).

Referring to (4.3) and (4.11), the functions Q_m and Q^* for the region R defined by

$$R : \bar{x} \leq \mu + a \quad \dots (5.51)$$

are given by
$$Q_m = (2\pi)^{-\frac{1}{2}r} \int \dots \int_{\xi \leq a\sqrt{m}/\sigma} e^{-\frac{1}{2}\xi'\xi} d\xi = \prod_1^r \Phi\left(\frac{a_i\sqrt{m}}{\sigma}\right), \quad \dots (5.52)$$

and
$$Q^* = \int_0^\infty \left\{ \prod_1^r \Phi(a_i k \sqrt{u}) \right\} dF_{v_0}(u), \quad \dots (5.53)$$

with $v_0 = n_0 q - r - s$ (as before). Consequently, v_0 and k must satisfy

$$\int_0^\infty \left\{ \prod_1^r \Phi(a_i k \sqrt{u}) \right\} dF_{v_0}(u) = 1 - \alpha. \quad \dots (5.54)$$

We note that Q^* is increasing in u , so that the confidence coefficient is certainly greater than $1 - \alpha$.

Since the a_i are arbitrary, equation (5.53) gives the joint distribution of the $\bar{x}_i - \mu_i$ ($i = 1, 2, \dots, r$). The joint distribution of the scaled random variables z_i , with $z_i = k(\bar{x}_i - \mu_i)$ is a form of the generalized t -distribution* (cf. (4.12) with (5.2)),

*See Dunnett and Sobel (1954, 1955).

TWO-STAGE SAMPLE MEANS FROM NORMAL POPULATIONS

the corresponding density function of which is readily obtained from (5.53) by differentiation with respect to the a_i as

$$\frac{\Gamma\left(\frac{\nu_0+r}{2}\right)}{(\nu_0\pi)^{1/2} \Gamma\left(\frac{\nu_0}{2}\right)} \left(1 + \frac{\sum_{i=1}^r z_i^2}{\nu_0}\right)^{-\frac{1}{2}(\nu_0+r)} \quad \dots \quad (5.55)$$

Joint one-sided tests for the μ_i with power independent of the variance may clearly be obtained in the same manner. Such tests are conservative in the sense discussed previously.

6. FURTHER DISCUSSION

(i) Suppose σ is regarded as having some prior distribution function $K(\sigma)$. Then

$$\begin{aligned} \Pr\{\bar{x} \in R | \mu, K\} &= \int_0^\infty P(R | \mu, \sigma) dK(\sigma) \geq P^*(R | \mu), P(R | \mu, \sigma) \text{ decreasing in } \sigma, \\ &\leq P^*(R | \mu), P(R | \mu, \sigma) \text{ increasing in } \sigma, \dots \quad (6.1) \end{aligned}$$

where $P(R | \mu, \sigma)$ and $P^*(R | \mu)$ are defined in (4.1) and (4.10), respectively. This means that the previous test procedures remain valid even if σ is not assumed to be constant but has some Bayesian distribution associated with it, provided $P(R | \mu, \sigma)$ is monotone in σ . The tests are still conservative in character if $P^*(R | \mu)$ is given as the power function.

If, further, both μ and σ are regarded as having a joint prior distribution $K(\mu, \sigma)$ and $P(R | \mu, \sigma)$ is independent of μ , $P(R | \mu, \sigma) = P(R | \sigma)$, then

$$\Pr\{\bar{x} \in R | K\} = \int_{-\infty}^\infty \int_0^\infty P(R | \sigma) dK(\mu, \sigma) \geq P^*(R), P(R | \sigma) \text{ decreasing in } \sigma \quad \dots \quad (6.2)$$

where
$$P^*(R) = \lim_{\sigma \rightarrow \infty} P(R | \sigma). \quad \dots \quad (6.3)$$

This implies that the previous estimation procedures for the μ_i , or of functions of the μ_i , by means of confidence regions likewise remain valid when μ and σ have prior distributions, provided $P(R | \sigma)$ is decreasing in σ . The estimation procedures are still conservative in character if $P^*(R)$ is given as the confidence coefficient.

(ii) The appropriate probabilities for the test and estimation procedures have been obtained previously by letting $\sigma \rightarrow \infty$, and the conservative character of the procedures deduced from considerations of monotonicity with respect to σ of the corresponding (conditional) probabilities for fixed total sample size. In practice,

however, one would almost always be in a position to set an upper bound to σ , say σ' (this in itself implies a weak specification for the prior distribution of σ , cf. (i)), and, therefore, to improve on the 'limiting' power or confidence coefficient attained asymptotically as $\sigma \rightarrow \infty$. As an example, the confidence coefficient for the case of a single mean and symmetrical confidence intervals of width $2a$ would be stated as at least

$$C(\sigma') = \sum_1^{\infty} p_m(\sigma') \left\{ 2\Phi \left(\frac{a\sqrt{m}}{\sigma'} \right) - 1 \right\}$$

(see (5.22)) rather than at least $1-\alpha$.*

(iii) Two-stage sampling procedures of the type discussed in this paper have been criticised in the past (Ruben, 1950; Lindley, 1958; Savage, 1959) on the general grounds of inefficiency, inasmuch as the observations from the second-stage sample are not used to supply information about σ . However, the loss in information can to some extent be recovered. As an example, consider again the case of symmetrical confidence estimation of a single mean. The mere conservative assertion that the confidence coefficient is greater than $1-\alpha$, though suitable for circumstances where the loss function is of such a character as to make the production of intervals with a covering probability for μ less than $1-\alpha$ unpleasant or undesirable or dangerous is nevertheless wasteful of information inasmuch as an infinitely high value of σ is assumed. This is merely another (though suggestive) aspect of the above criticism that not all the observations are used in assessing the variability of the population. Nevertheless, in principle there is nothing to prevent the statistician using all the observations to estimate σ^2 and thereby obtaining a better notion of the covering probability $C(\sigma)$. Thus, if s is the estimate of σ based on all the observations, calculated in the usual way, one could rather loosely say that the true covering probability is *about* $C(s)$, and a somewhat more comprehensive, if rather imprecise, assertion could be obtained by setting an upper bound, in a statistical sense, for σ based on s . For instance, having computed s one could then proceed to compute $s' = s + 1.48$ (standard error of s), the standard error of s being derived *approximately* by regarding n as fixed and so computed by the usual large sample formula. To illustrate by a numerical example, suppose the $C(s') = 0.90$ and $\alpha = 0.05$. Since the value s' corresponds to the ordinate in the approximate distribution of s which cuts off a probability of 0.10 to the right, one could then assert that in addition to the true covering probability being with absolute certainty greater than 0.95 the odds are 9 to 1 that the true covering probability is not less than 0.90. In brief, there is no real reason why the data should be 'reduced' to the extent that only a *single* number, viz. the confidence coefficient, is supplied as a measure of reliability of the estimate of the mean. Reduction or summarisation of data may frequently be unrealistic inasmuch as it artificially narrows

*We assume here that $C(\sigma)$ is strictly decreasing in σ (see footnote on p. 7), and numerical work seems to confirm this assumption.

TWO-STAGE SAMPLE MEANS FROM NORMAL POPULATIONS

the interpretative scope of a body of data. In our case, it appears to be more realistic to supply an ancillary statement of the type just mentioned which shall incorporate the additional information relevant to σ supplied by the additional observations in the second sample, or perhaps better still one could supply a graph or table, and ask any interested party to draw his own conclusions or make his own decisions contingent on any particular sample.

We may approach the matter from yet another angle. It has been shown previously¹ that the efficiency of the procedure discussed here in relation to the optimum fixed sample size procedure is $(t_{\infty; \alpha/2/t_{n_0}-1; \alpha/2})^2$ when σ^2 is large.² The efficiency is therefore quite high for large σ even if n_0 is only moderately large. On the other hand, if σ is small s is also likely to be small and, concomitantly, $C(s)$ is likely to be large (i.e. near to 1); and this will be so even if s_0^2 , and consequently n , are abnormally large, the high value of $C(s)$ thus to some extent compensating for the large amount of sampling.

We emphasize that this discussion is somewhat imprecise (better methods of recovering the lost information about σ may, and almost certainly do, exist), but this fault appears to be part of the price that one must be prepared to pay for having to meet the rather difficult (and perhaps artificial) objective of deriving estimation intervals of *fixed* length and being at the same time restricted to only two-stage rather than fully sequential sampling. The situation would naturally be different if it were desired to control the width of the confidence intervals, again of preassigned confidence coefficient, not absolutely but only stochastically, e.g. by imposing some such condition that the length of an interval exceeding a specified value be not greater than some specified (small) upper bound. In that situation, one would not need to have recourse to any such circuitous manner of recovering information about σ^2 as described above, since the predetermined confidence coefficient objective together with the ancillary objective of the stochastic control of the confidence intervals between them do not impose such drastic requirements on the reduction of data as does a mere specification of confidence coefficient and length of the confidence intervals. One could then use all the observations for the estimation of σ^2 more directly.

REFERENCES

- BARNARD, G. A. (1946): Sequential tests in industrial statistics. *Supp. J. Roy. Stat. Soc.*, **8**, 1-21.
 DANTZIG, G. B. (1940): On the non-existence of tests of Student's hypothesis having power functions independent of σ . *Ann. Math. Stat.*, **11**, 186-192.
 DUNNETT, C. W. and SOBEL, M. (1954): A bivariate generalisation of Student's t -distribution, with tables for certain special cases. *Biometrika*, **41**, 153-169.

¹Obtained as a particular case of (5.36). (See also Stein, 1945).

²More precisely, $k^2\sigma^2 > n_0$, as stated after (3.5). Note that by (5.16) the latter inequality is equivalent to $\frac{\sigma}{\sqrt{n_0}}/a > 1/t_{n_0-1; \alpha/2}$, i.e. the ratio of the standard error of the pilot sample to the precision of the sequentially estimated mean should be not small in relation to $1/t_{n_0-1; \alpha/2}$.

SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS : SERIES A

- DUNNET, C. W. and SOBEL, M. (1955): Approximations to the probability integral and certain percentage points of a multivariate analogue of Student's t -distribution. *Biometrika*, **42**, 259-260.
- LINDLEY, D. V. (1958): Professor Hogben's 'crisis'—a survey of the foundations of statistics. *Applied Statistics*, **7**, 186-198.
- PATNAIK, P. B. (1949): The non-central χ^2 and F -distributions and their application. *Biometrika*, **25**, 202-232.
- RUBEN, H. (1950): *Sequential Studentisation of Sampling Means from Unknown Normal Populations*. Ph. D. dissertation, London University.
- SAVAGE, L. J. (1959): Remarks at conference on statistical inference held at Birkbeck College, London University, July, 27-29. (To be published).
- SEELBINDER, B. M. (1953): On Stein's two-stage sampling scheme. *Ann. Math. Stat.*, **24**, 640-649.
- STEIN, C. (1945): A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Stat.*, **16**, 243-258.
- WALD, A. (1947): *Sequential Analysis*, John Wiley and Sons.

Paper received: March, 1960.

TESTING FOR THE SINGLE OUTLIER IN A REGRESSION MODEL

By K. S. SRIKANTAN*

Indian Statistical Institute

SUMMARY. The general problem of testing a regression model against an alternative hypothesis of a single outlier is treated in this paper. Certain test criteria are developed and their 'nominal percentage points', which control the significance of the tests at or below prescribed levels, are obtained.

INTRODUCTION

An experimenter may be confronted with the hypothesis that the observations he has obtained are independent and normal with a linear regression on a known set of variables. He may contemplate an alternative in which one or a specified number of observations, *though he does not know precisely which*, can deviate from the assumed regression or can have increased variability. Here we have to test the hypothesis of the given regression against the alternative of a specified number of 'outliers'. The test should be formulated in such a manner that it is 'best' in detecting any of the alternatives when they happen to be true and, at the same time, controls the error of rejecting the hypothesis, when true, at a preassigned level of probability. This is the problem of testing for outliers.

This problem differs from the analysis of variance. In the latter the regression model is tested against all possible alternatives of shift in the means of the observations. But in the outlier problem the alternatives are restricted to the deviation in the mean of one or a specified number of any of the observations from the given type of regression.

It is seen, therefore, that the outlier test is appropriate to a situation in which we possess the additional information (over and above that required for the analysis of variance) that only a certain number of any of the observations can deviate, in their mean, from the assumed regression. In some problems such information is forthcoming. For example, in factories, it might be found, from past experience, that at a time only one or two machines went out of alignment.

When the number of outliers contemplated in the alternative hypothesis is just one, the test is called a single outlier test. This paper deals with such tests.

The problem of testing for a single outlier when the hypothesis is that all the observations come from identical and independent normal populations with the same but unknown mean and variance has been completely solved. The first significant contribution towards this problem was made by Pearson and Chandra Sekhar (1936) based on the work of Thompson (1935). They suggested certain test criteria and also obtained their percentage points in small samples. Later on percentage

* Now with Central Statistical Organisation, New Delhi.

points for these criteria were obtained for larger sample sizes also by Grubbs (1950). Percentage points for a slightly different criterion involving an external estimate of variance were obtained by a different method by Nair (1948).

The present paper deals with the more general problem of the single outlier in a regression model. This, no doubt, is a more difficult problem. In this case, the distributions of the test criteria, apart from being difficult to solve, should involve, when solved, as parameters, certain functions of the known variables on which the observations have a linear regression. In order to avoid this only the 'nominal percentage points' of these criteria have been obtained. The 'nominal percentage points' always control the error of the first kind at a level not exceeding the specified one; and under some conditions, which hold in restricted cases in small samples, these points control the error of the first kind actually at the specified level. 'Nominal percentage points' are used in this special sense throughout this paper.

The method followed in this paper, for obtaining the nominal percentage points, would be seen to resemble in some respects that adopted by Pearson and Chandra Sekhar (1936) and in certain others that of Cochran (1941). In the next section we formulate the problem of the single outlier and the criteria for its detection.

2. SINGLE OUTLIER IN A REGRESSION MODEL

2.1. *Preliminaries.* Let $\underline{y} = (y_1, y_2, \dots, y_n)$ be n independently and normally distributed random variables with the same unknown variance σ^2 and means to be specified later according to the alternative hypotheses considered. Let $\underline{\beta} = (\beta_1, \beta_2, \dots, \beta_m)$ be m unknown parameters and $X = \{x_{i,j}\}$ be a matrix of $m \times n$ known constants of rank m where $m < n$. Further let

$$\mu_j = \beta_1 x_{1,j} + \beta_2 x_{2,j} + \dots + \beta_m x_{m,j}. \quad \dots (2.1)$$

If $E(y_j) = \mu_j$, $j = 1(1)n$, then the least square estimate of $\underline{\beta}$ is given by

$$\underline{b} = \underline{y}.X'.(X.X')^{-1}. \quad \dots (2.2)$$

We shall represent the deviations of the observed values from their regression estimates by the vector

$$\underline{e} = (e_1, e_2, \dots, e_n);$$

thus

$$\underline{e} = \underline{y} - \underline{b}.X. \quad \dots (2.3)$$

It could be shown that the dispersion matrix of $\underline{e}$ is $\Lambda\sigma^2$ where Λ is the $n \times n$ matrix

$$\{\lambda_{i,j}\} = I - X'(X.X')^{-1}.X. \quad \dots (2.4)$$

The error sum of squares will reduce to

$$S^2 = \underline{e} . \underline{e}' = \underline{y} . \underline{\Lambda} . \underline{y}'. \quad \dots (2.5)$$

TESTING FOR THE SINGLE OUTLIER IN A REGRESSION MODEL

We shall use the following statistics:

$$d_i = e_i \{\lambda_{i,i} S^2\}^{-\frac{1}{2}}, \quad \dots (2.6)$$

$$t_i = d_i^2, \quad \dots (2.7)$$

$$\begin{aligned} u_i &= d_i^2 \quad \text{if } d_i \geq 0 \\ &= 0 \quad \text{if } d_i < 0 \end{aligned} \quad \dots (2.8)$$

$$\begin{aligned} l_i &= 0 \quad \text{if } d_i \geq 0 \\ &= d_i^2 \quad \text{if } d_i < 0. \end{aligned} \quad \dots (2.9)$$

It is well known from the general theory of least squares that d_i is symmetrically distributed and that t_i follows the Beta distribution with parameters

$$(\frac{1}{2}, (n-m-1)/2).$$

Consequently, for any $x > 0$ we have

$$\text{Prob } (u_i \geq x) = \text{Prob } (d_i \geq \sqrt{x}) = \frac{1}{2} - \frac{1}{2} I_x(\frac{1}{2}, (n-m-1)/2) \quad \dots (2.10)$$

$$\text{where} \quad I_x(p, q) = \left\{ \int_0^x \xi^{p-1} (1-\xi)^{q-1} d\xi \right\} / B(p, q) \quad \dots (2.11)$$

Similarly for $x > 0$,

$$\text{Prob } (l_i \geq x) = \text{Prob } (d_i \geq -\sqrt{x}) = \frac{1}{2} - \frac{1}{2} I_x(\frac{1}{2}, (n-m-1)/2). \quad \dots (2.12)$$

2.2. *Test for a single specified outlier.* When it is specified that the i -th observation is an outlier, we can set up

$$E(y_j) = \mu_j, \quad j \neq i,$$

$$\text{and} \quad E(y_i) = \mu_i + \delta_i \quad (i \text{ specified and } \mu\text{'s given by (2.1)}),$$

and express the null hypothesis H_0 and the alternative hypothesis H in terms of δ_i as follows:

$$H_0 : \delta_i = 0$$

$$\text{and} \quad H : \delta_i = \delta \neq 0.$$

Let k_1 be the upper $200\alpha\%$ point and k_2 the upper $100\alpha\%$ point of the distribution of t_i . It could be easily shown that against one sided alternatives, uniformly most powerful similar regions exist and, corresponding to the level of significance α , these are given by

$$u_i \geq k_1 \quad \text{if } \delta > 0 \quad \dots (2.13)$$

and

$$l_i \geq k_1 \quad \text{if } \delta < 0. \quad \dots (2.14)$$

Against both sided alternatives, the best unbiased test, which controls the error of the first kind at the level α , is

$$t_i \geq k_2. \quad \dots (2.15)$$

2.3. *Test for a single unspecified outlier.* Let

$$t = \max(t_1, t_2, \dots, t_n) = \max_i e_i^2 / \lambda_{i,i} \quad y \wedge y' \quad \dots (2.16)$$

$$u = \max(u_1, u_2, \dots, u_n) = \max_i e_i |e_i| / \lambda_{i,i} \quad y \wedge y' \quad \dots (2.17)$$

and

$$l = \max(l_1, l_2, \dots, l_n) = \max_i -e_i / |e_i| \lambda_{i,i} \quad y \wedge y' \quad \dots (2.18)$$

where λ 's are defined by (2.4) and e 's by (2.2) and (2.3).

When the outlier is not specified, the results of Section 2.2 immediately suggest the use of one of the criteria t , u or l , these having the common property of being the maximum of the studentised squared deviations in some appropriate sense. The criterion u should be used when the alternative hypothesis is that the expected value of the outlier exceeds that given by the regression model; l should be used against one sided alternatives in the other direction and t against both sided alternatives. Though from intuitive considerations these tests seem to be justifiable, their performance will not be examined in this paper.

The problem investigated here is the evaluation of the percentage points of these criteria. This will be discussed in the next section.

3. EVALUATION OF THE PERCENTAGE POINTS OF THE DISTRIBUTION OF A MAXIMUM

3.1. *Nominal percentage points.* Let $(v_1, v_2, \dots, v_n)$ be n random variables and further let $v = \max(v_1, v_2, \dots, v_n)$. Also let

$$P(V) = \text{Prob}(v \geq V) \quad \dots (3.1)$$

$$P_1(V) = \sum_{i=1}^n \text{Prob}(v_i \geq V) \quad \dots (3.2)$$

and

$$P_2(V) = \sum_{i=2}^n \sum_{j=1}^{i-1} \text{Prob}(v_i \geq V, v_j \geq V). \quad \dots (3.3)$$

TESTING FOR THE SINGLE OUTLIER IN A REGRESSION MODEL

Then we could derive from the probability laws for the joint occurrence of n events, that

$$P_1(V) - P_2(V) \leq P(V) \leq P_1(V). \quad \dots (3.4)$$

If V_0 is the exact upper 100 α % point of the distribution of v , then

$$P(V_0) = \alpha. \quad \dots (3.5)$$

The quantity V_1 defined by $P_1(V_1) = \alpha$... (3.6)

will be called the nominal upper 100 α % point of the distribution of v . From (3.4) it follows that

$$P(V_1) \leq \alpha. \quad \dots (3.7)$$

Thus, in testing any alternative hypothesis against the upper tail determined by the nominal percentage point V_1 , the actual level of significance cannot exceed α . Further if $P_2(V_1)$ happens to be zero then the nominal percentage point coincides with the actual one; a sufficient condition for this is:

$$\text{Prob } (v_i \geq V_1, v_j \geq V_1) = 0 \text{ for all } i \neq j. \quad \dots (3.8)$$

For the test criteria developed in Section 2.3, the actual percentage points are difficult to evaluate and would, in general, depend on certain parametric functions of the unknown variables (x 's) on which the observations have a linear regression. In order to avoid this only nominal percentage points will be obtained in this paper. However, for special types of regression functions, it will be shown that for samples of sizes below a specified number, these nominal percentage points turn out to be the actual ones. For this purpose we require the two lemmas proved below.

3.2. *Two useful lemmas.* Let Z_1, Z_2 and W be three real valued random variables such that

$$\text{Prob } (W \leq 0 | Z_1, Z_2) = 0 \text{ for all } Z_1 \text{ and } Z_2. \quad \dots (3.9)$$

Let ρ be a real number where $|\rho| < 1$ and let

$$Q^2 = \{(Z_1^2 - 2\rho Z_1 Z_2 + Z_2^2)/(1 - \rho^2)\} + W \quad \dots (3.10)$$

and $R_i = Z_i/Q$ ($i = 1, 2$). Further let h be any given positive number. Then we have the following.

Lemma 1: $\text{Prob } (R_1 \geq h, R_2 \geq h) = 0$ if $2h^2 \geq 1 + \rho$ and

Lemma 2: $\text{Prob } (|R_1| \geq h, |R_2| \geq h) = 0$ if $2h^2 \geq 1 + |\rho|$.

Proof: Obviously,

$$U = R_1^2 - 2\rho R_1 R_2 + R_2^2 < 1 - \rho^2 \text{ by (3.9) and (3.10).}$$

But
$$U = (R_1 - R_2)^2 + 2R_1R_2(1 - \rho).$$

Hence when
$$R_1 \geq h \text{ and } R_2 \geq h,$$

$$U \geq 2h^2(1 - \rho) \geq 1 - \rho^2 \text{ if } 2h^2 \geq 1 + \rho.$$

Therefore we have Lemma 1.

Again,
$$U \geq |R_1|^2 - 2|\rho||R_1||R_2| + |R_2|^2 = (|R_1| - |R_2|)^2 + 2|R_1||R_2|(1 - |\rho|).$$

Thus when
$$|R_1| \geq h \text{ and } |R_2| \geq h,$$

$$U \geq 2h^2(1 - |\rho|) \geq 1 - \rho^2 \text{ if } 2h^2 \geq 1 + |\rho|$$

and Lemma 2 follows.

3.3. *Evaluation of nominal percentage points of u and l .* Since $u = \max(u_1, u_2, \dots, u_n)$ it follows from (3.6) and (2.10) that the nominal upper $100\alpha\%$ point of u is given by the following equation in x :

$$(n/2) \cdot [1 - I_x(\frac{1}{2}, (n-m-1)/2)] = \alpha. \quad \dots (3.11)$$

It also follows from (3.8) that a sufficient condition for this nominal percentage point to coincide with the actual one is that

$$\text{Prob}(u_i \geq x, u_j \geq x) = 0 \text{ for all } i \neq j,$$

that is,
$$\text{Prob}(d_i \geq \sqrt{x}, d_j \geq \sqrt{x}) = 0.$$

By an appropriate real linear transformation from

$$(y_1, y_2, \dots, y_n) \text{ to } (z_1, z_2, \dots, z_n),$$

d_i and d_j could be represented as

$$d_i = z_1/S \text{ and } d_j = z_2/S$$

where
$$S^2 = \{(z_1^2 - 2\rho_{i,j}z_1z_2 + z_2^2)/(1 - \rho_{i,j}^2)\} + z_3^2 + z_4^2 + \dots + z_{n-m}^2 \quad \dots (3.12)$$

with $\rho_{i,j} = \lambda_{i,j} \{\lambda_{i,i} \lambda_{j,j}\}^{-\frac{1}{2}}$ (λ 's being defined by (2.4)), and $\text{Prob}(S \leq 0/z_1, z_2) = 0$ for all z_1 and z_2 . Applying Lemma 1 we get the following necessary condition for the nominal percentage point x to coincide with the actual one :

$$2x \geq 1 + \rho_{i,j} = 1 + \lambda_{i,j} \{\lambda_{i,i} \lambda_{j,j}\}^{-\frac{1}{2}} \quad \dots (3.13)$$

(λ 's being defined by (2.4)).

The nominal percentage points of l turn out to be the same as those of u in virtue of (2.12); (3.13) becomes a sufficient condition for the coincidence of the nominal and actual percentage points of l .

TESTING FOR THE SINGLE OUTLIER IN A REGRESSION MODEL

3.4. *Evaluation of the nominal percentage points of t .* Since $t = \max(t_1, t_2, \dots, t_n)$ where t_i follows the B distribution with parameters $(\frac{1}{2}, (n-m-1)/2)$ (Section 2.1), it is seen from (3.6) that the nominal upper $100\alpha\%$ point of the distribution of t is given by the equation in x :

$$n[1 - I_x(\frac{1}{2}, (n-m-1)/2)] = \alpha. \quad \dots (3.14)$$

A sufficient condition for x to be the actual percentage point is that

$$\text{Prob } (t_i \geq x, t_j \geq x) = 0 \text{ for all } i \neq j;$$

that is,

$$\text{Prob } (|d_i| \geq \sqrt{x}, |d_j| \geq \sqrt{x}) = 0.$$

Proceeding as in Section 3.3 and using Lemma 2, we find that a sufficient condition for this is

$$2x \geq 1 + |\rho_{i,j}| = 1 + |\lambda_{i,j}| \{\lambda_{i,i} \lambda_{j,j}\}^{-\frac{1}{2}} \quad \dots (3.15)$$

where λ 's are defined by (2.4).

4. SCOPE OF THE TABLES OF NOMINAL PERCENTAGE POINTS

Tables of the nominal upper 5% and 1% points of the distribution of the criteria u (or l) and t are presented in Tables 1 to 3 for sample sizes up to $n = 20$ and regression variables $m = 1, 2$ and 3. These were computed by solving equations (3.11) and (3.14) for $\alpha = 0.05$ and 0.01. For this purpose Karl Pearson's Incomplete Beta Function Tables were used and Newton's divided difference formula was applied to carry out inverse interpolation of the second order. The results are expected to be correct to within one digit in the fourth decimal place.

5. SPECIAL REGRESSIONS: COINCIDENCE OF NOMINAL AND ACTUAL PERCENTAGE POINTS

5.1. *Regression on one variable ($m = 1$).* For the simple case $E(y_j) = \beta$, $j = 1(1)n$, the test criteria u and l reduce to those formulated by Pearson and Chandra Sekhar (1936) and Grubbs (1950). We have, in fact,

$$u = 1 - S_n^2 / S^2 \quad (\text{Grubbs}) \quad \dots (5.1)$$

$$= \{T^{(n)}\}^2 / (n-1) \quad (\text{Pearson and Chandra Sekhar}) \quad \dots (5.2)$$

$$\text{and } l = 1 - S_1^2 / S^2 \quad (\text{Grubbs}) \quad \dots (5.3)$$

$$= \{T^{(1)}\}^2 / (n-1) \quad (\text{Pearson and Chandra Sekhar}). \quad \dots (5.4)$$

It was found that the nominal percentage points of u and l agreed to all four places of decimals with the actual ones obtained by Grubbs.

Since, in this case,

$$\rho_{i,j} = (-1)/(n-1) \quad \text{for all } i \neq j,$$

the nominal 5% points of u (or l) are actual up to sample size 14 and the 1% points up to 19. The nominal 5% points of t are up to sample size 13 and 1% points up to 18.

5.2. *Regression on two variables* ($m = 2$). An important special regression is

$$E(y_j) = \beta_1 + \beta_2 \cdot j, \quad j = 1(1)n. \quad \dots (5.5)$$

In this case we get after simplification,

$$\Lambda = I - \{\alpha_{i,j}\} \quad \dots (5.8)$$

where $\alpha_{i,j} = \{(n^2-1) + 12(i-[n+1]/2)(j-[n+1]/2)\}/n(n^2-1)$. Therefore, $\max \rho_{i,j} = 2/(n-1)$. Thus for this regression the nominal 5% points of u (or l) are actual up to sample size 10, and the 1% points are actual up to 15. Again,

$$\max |\rho_{i,j}| = 4\{n^2 - 2n + 13\}^{-1/2}.$$

Therefore, the nominal 5% points of t are actual up to sample size 9 and the 1% points up to 14.

5.3. *Regression on three variables* ($m = 3$). The special regression considered is

$$E(y_j) = \beta_1 + \beta_2 \sin(2\pi j/\lambda) + \beta_3 \cos(2\pi j/\lambda) \quad \dots (5.7)$$

where λ is a known positive integer and $j = 1(1)n$.

When n is a multiple of λ , we have

$$\Lambda = I - \{\alpha_{i,j}\} \quad \dots (5.8)$$

where

$$\alpha_{i,j} = \{1 + 2 \cos[2\pi(i-j)/\lambda]\}/n.$$

Thus

$$\rho_{i,j} = -\{1 + 2 \cos[2\pi(i-j)/\lambda]\}/(n-3) \leq 1/(n-3). \quad \dots (5.9)$$

So the nominal 5% points of u (or l) are actual up to sample size 13 and the 1% points up to 18 for this regression.

ACKNOWLEDGEMENT

The author acknowledges, with gratitude, his debt to Dr. P. B. Patnaik of the Indian Statistical Institute for his kind encouragement and useful suggestions in the preparation of this paper; to Dr. Des Raj of the same Institute who proposed the problem of the single outlier; to Shri Sivaramakrishnan who helped in the computations; and to the referee of this paper who suggested many improvements and made the presentation more lucid.

TESTING FOR THE SINGLE OUTLIER IN A REGRESSION MODEL

NOMINAL UPPER PERCENTAGE POINTS IN UNITS OF 10^{-4}

TABLE 1*

Regression : $E(y_j) = \beta_1 x_{1,j}$, $j = 1(1)n$

TABLE 2

$E(y_j) = \beta_1 x_{1,j} + \beta_2 x_{2,j}$, $j = 1(1)n$

TABLE 3

$E(y_j) = \beta_1 x_{1,j} + \beta_2 x_{2,j} + \beta_3 x_{3,j}$, $j = 1(1)n$

sample size	u or l		t		sample size	u or l		t		sample size	u or l	
	5%	1%	5%	1%		5%	1%	5%	1%		5%	1%
(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)
3	9973	9998	9993	10000	4	9985	9999	9996	10000	5	9993	10000
4	9506	9900	9752	9950	5	9604	9920	9801	9960			
5	8730	9558	9192	9721								
6	7968	9072	8547	9340	6	8872	9608	9283	9753	6	9669	9933
7	7304	8553	7934	8897	7	8114	9140	8652	9389	7	8979	9645
8	6739	8052	7384	8446	8	7438	8626	8038	8954	8	8231	9195
9	6258	7589	6899	8011	9	6858	8125	7481**	8504	9	7551	8688
10	5846	7169	6474	7606	10	6363**	7660	6987	8069	10	6961	8188
11	5489	6789	6099	7233	11	5938	7233	6553	7661	11	6455	7720
12	5178	6446	5768	6890	12	5570	6848	6170	7284	12	6020	7291
13	4903	6136	5472**	6576	13	5249	6500	5831	6938	13	5644**	6901
14	4660	5855	5208	6289	14	4967	6185	5530	6621**	14	5315	6549
15	4441	5599	4970	6026	15	4716	5900**	5259	6330	15	5026	6231
16	4245	5366	4754	5784	16	4493	5640	5011	6064	16	4769	5942
17	4067	5152	4558	5561	17	4291	5404	4797	5820	17	4540	5679
18	3905	4956	4379	5356**	18	4109	5187	4597	5594	18	4334	5439**
19	3757	4775	4215	5165	19	3943	4988	4416	5386	19	4149	5220
20	3621	4607	4063	4989	20	3792	4805	4248	5094	20	3979	5018

*Table 1 provides actual percentage points of u and l for the regression $E(y_j) = \beta$ (Section 5.1).

**Percentage points are actual up to and including this sample size for the regression

Table 1 : $E(y_j) = \beta$; Table 2 : $E(y_j) = \beta_1 + \beta_2 \cdot j$ (Section 5.2); Table 3 : $E(y_j) = \beta_1 + \beta_2 \sin(2\pi j/\lambda) + \beta_3 \cos(2\pi j/\lambda)$, $j = 1(1)n$, where λ is a positive integer and n is a multiple of λ (Section 5.3).

Note : 1. u , l and t are given by equations (2.16), (2.17) and (2.18).

2. Nominal percentage points of u and l satisfy (3.11) and those of t satisfy (3.14).

3. Condition for coincidence of the nominal and actual percentage points of u and l is given by (3.13) and for t by (3.15).

4. In virtue of (3.13) and (3.15) the nominal upper 100 α percent points of u (or l) are the same as the 200 α percent points of t . Thus the nominal 1% and 5% points of u could be used as the 2% and 10% points of t and nominal 1% and 5% points of t could be used as 0.5% and 2.5% points of u (or l).

REFERENCES

- CHANDRA SEKHAR, C. and PEARSON, E. S. (1936): The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, **28**, 308-320.
- COCHRAN, W. G. (1941): The distribution of the largest of a set of estimated variances as a fraction of their total. *Ann. Eugen.*, **II**, 47-52.
- GODWIN, H. J. (1945): On the distribution of the estimate of mean deviation obtained from samples from a normal population. *Biometrika*, **33**, 254-256.
- GRUBBS, F. E. (1950): Testing outlying observations. *Ann. Math. Stat.*, **21**, 27-58.
- HARTLEY, H. O. and PEARSON, E. S. (1942): The probability integral of the range in samples of n observations from a normal population. *Biometrika*, **32**, 301-310.
- (1943): Tables of the probability integral of the studentized range. *Biometrika*, **33**, 89-99.
- IRWIN, J. O. (1925): On a criterion for the rejection of outlying observations. *Biometrika*, **17**, 233-250.

SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS : SERIES A

- MCKAY, A. T. (1935): The distribution of the difference between the extreme observation and the sample mean in samples of n from a normal universe. *Biometrika*, **27**, 466-471.
- NAIR, K. R. (1948): The distribution of the extreme deviate from the sample mean and its studentized form. *Biometrika*, **35**, 118-144.
- NEYMAN, J. and PEARSON, E. S. (1933): On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc., (Lond.)* **231**, 289-337.
- PEARSON, E. S. (1925): A further note on the distribution of range in samples taken from a normal population. *Biometrika*, **17**, 151-164.
- PEARSON, K. (1934): *Tables of the Incomplete Beta-Function*. Biometrika Office, University College, London.
- (1931): *Tables for Statisticians and Biometricians*, Part II, edited by Karl Pearson, CX-CXIX.
- SMIRNOFF, N. V. (1941): On the estimation of the maximum term in a series of observations. C. R. (Doklady). *Acad. Sci. (N.S.)*, **33** 346-50.
- "Student", (1927): *Biometrika*, **19**, 151-164.
- THOMPSON, W. R. (1935): On a criterion for the rejection of observations and the distribution of the ratio of the deviation to the sample standard deviation. *Ann. Math. Stat.*, **6**, 214-219.
- TIPPET, L. H. C. (1925): The extreme individuals and the range of samples taken from a normal population. *Biometrika*, **17**, 151-164.
- W. P. A. (1945): Tables of probability integral of the mean deviation in normal samples. *Biometrika*, **33**, 259-265.

Paper received: March, 1956.

Revised: October, 1960.

MINIMAL SUFFICIENT STATISTICS FOR THE BALANCED INCOMPLETE BLOCK DESIGN UNDER AN EISENHART MODEL II*

By DAVID L. WEEKS

and

FRANKLIN A. GRAYBILL

Oklahoma State University

SUMMARY. The authors present in this paper a set of minimal sufficient statistics for a General Balanced Incomplete Block model under the assumptions that the treatments and the errors are random variables, i.e., Eisenhart Model II.

1. INTRODUCTION

Sufficient statistics play a very important role in the theory of minimum variance unbiased estimation [(Rao, 1946); (Blackwell, 1947)]. If a minimal set of sufficient statistics can be found for a particular distribution, then the Rao-Blackwell theorem states that minimum variance unbiased estimators must be an explicit function of this set. If the minimal set is also complete (Lehmann and Scheffé, 1950), then an unbiased estimator based on this complete set is the *unique minimum variance unbiased estimator*. In confidence interval estimation and tests of hypotheses minimal sufficient statistics play a similar role.

The distribution of a set of minimal sufficient statistics can be considered as the canonical form of the distribution of the original observations since this set represents the maximum reduction possible without losing "information." Therefore, it is quite useful to be able to exhibit a set of minimal sufficient statistics and their joint distribution.

The purpose of this paper is to exhibit a set of minimal sufficient statistics for a general balanced incomplete block model under the assumptions that the blocks, the treatments, and the errors are random variables; i.e., Eisenhart's Model II (Eisenhart, 1947)

2. DEFINITIONS AND ASSUMPTIONS

A balanced incomplete block is defined as a design in which there are t treatments and b blocks of k experimental units per block (where $k < t$) with each treatment replicated r times. The arrangement of treatments is such that every pair of treatments occurs together in exactly λ blocks.

* Research sponsored in part by National Science Foundation, Grant No. NSF-G-3970.

The model may be written as a special case of the general two-way classification (Kempthorne, 1952), i.e.,

$$Y_{ijm} = \mu + \beta_i + \tau_j + e_{ijm} \quad \dots (2.1)$$

where $i = 1, 2, \dots, b$; $j = 1, 2, \dots, t$; $m = n_{ij}$; where

$$n_{ij} = \begin{cases} 1 & \text{if treatment } j \text{ occurs in block } i. \\ 0 & \text{otherwise.} \end{cases}$$

It should be noted that only the Y_{ijm} in which $m \neq 0$ are observed.

Equation (2.1) represents $bk=n$ equations and these equations may be written in matrix form as

$$\mathbf{Y} = \mu \mathbf{X}_0 + \mathbf{X}_1 \boldsymbol{\beta} + \mathbf{X}_2 \boldsymbol{\tau} + \mathbf{e} \quad \dots (2.2)$$

where the dimensions of the matrices in (2.2) are as follows: $\mathbf{Y}(n \times 1)$; $\mathbf{X}_1(n \times b)$; $\boldsymbol{\beta}(b \times 1)$; $\mathbf{X}_2(n \times t)$; $\boldsymbol{\tau}(t \times 1)$; $\mathbf{e}(n \times 1)$; $\mu(1 \times 1)$; with $\mathbf{X}_0$ an $n \times 1$ vector with each element equal to one. ($\mathbf{j}_1^s$ will be used to denote an $s \times 1$ vector of ones and $\mathbf{J}_q^s$ an $s \times q$ matrix of ones). We will denote $\mathbf{X}_2' \mathbf{X}_1$ by $\mathbf{N}$ and the $q \times q$ identity matrix by $\mathbf{I}_q$.

Under an Eisenhart Model II we shall assume the following distributional properties of the vectors $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, and $\mathbf{e}$.

(a) $\boldsymbol{\beta}$ is distributed as the multivariate normal, mean $\mathbf{0}$, covariance matrix $\sigma_1^2 \mathbf{I}_b$.

(b) $\boldsymbol{\tau}$ is distributed as the multivariate normal, mean $\mathbf{0}$, covariance matrix $\sigma_2^2 \mathbf{I}_t$.

(c) $\mathbf{e}$ is distributed as the multivariate normal, mean $\mathbf{0}$, covariance matrix $\sigma^2 \mathbf{I}_n$.

(d) $\text{cov}(\boldsymbol{\beta}, \boldsymbol{\tau}) = \mathbf{0}$, $\text{cov}(\boldsymbol{\beta}, \mathbf{e}) = \mathbf{0}$, $\text{cov}(\boldsymbol{\tau}, \mathbf{e}) = \mathbf{0}$.

(e) μ is a scalar constant.

Under these assumptions, the joint distribution of the elements of the $\mathbf{Y}$ is given by

$$g(\mathbf{Y}; \boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2} |\mathbf{V}|^{1/2}} \exp -\frac{1}{2} (\mathbf{Y} - \bar{\boldsymbol{\mu}})' \mathbf{V}^{-1} (\mathbf{Y} - \bar{\boldsymbol{\mu}}) \quad \dots (2.3)$$

where

$$\bar{\boldsymbol{\mu}} = \mu \mathbf{X}_0, \boldsymbol{\theta}' = (\mu, \sigma^2, \sigma_1^2, \sigma_2^2)$$

$$\mathbf{V} = (\mathbf{X}_1 \mathbf{X}_1' \sigma_1^2 + \mathbf{X}_2 \mathbf{X}_2' \sigma_2^2 + \mathbf{I}_n \sigma^2).$$

The following relationships hold for the matrix model :

$$(a) \quad \mathbf{X}_1' \mathbf{X}_1 = k \mathbf{I}_b$$

$$(d) \quad \mathbf{J}_b^n \mathbf{X}_1' = \mathbf{J}_n^n$$

$$(g) \quad \mathbf{N} \mathbf{N}' = (r - \lambda) \mathbf{I}_t + \lambda \mathbf{J}_t^t$$

$$(b) \quad \mathbf{X}_2' \mathbf{X}_2 = r \mathbf{I}_t$$

$$(e) \quad \mathbf{J}_n^n \mathbf{X}_2' = r \mathbf{J}_t^n$$

$$(h) \quad (\mathbf{X}_2' - k^{-1} \mathbf{N} \mathbf{X}_1') \mathbf{X}_2 = \lambda k^{-1} (t \mathbf{I}_t - \mathbf{J}_t^t)$$

$$(c) \quad \mathbf{J}_n^n \mathbf{X}_1 = k \mathbf{J}_b^n$$

$$(f) \quad \mathbf{J}_t^n \mathbf{X}_2' = \mathbf{J}_n^n$$

$$(i) \quad (\mathbf{X}_2' - k^{-1} \mu \mathbf{X}_1') \mathbf{X}_1 = \mathbf{0}.$$

3. DEVELOPMENT OF A SET OF MINIMAL SUFFICIENT STATISTICS

In this section we will define an orthogonal matrix P in such a way that by transforming Y through the transformation P , a set of sufficient statistics will be obtained. Let P_3 be a set of $t-1$ orthogonal vectors such that

$$P_3' N N' P_3 = D_3$$

where D_3 is diagonal with $t-1$ of the characteristic roots of NN' on the diagonal. The characteristic roots of NN' are rk and $(r-\lambda)$ of multiplicities 1 and $t-1$, respectively. Let P_3^* be constructed such that

$$P_3^* N N' P_3^* = \begin{bmatrix} rk & 0 \\ 0 & (r-\lambda)I_{t-1} \end{bmatrix}.$$

Let the first row of P_3^* be $\frac{1}{\sqrt{t}} j_t^1$ and P_3' be a set of $t-1$ orthogonal vectors which are

orthogonal to $\frac{1}{\sqrt{t}} j_t^1$. Hence

$$\frac{1}{\sqrt{t}} j_t^1 P_3' = 0.$$

Then from the form of NN' in the BIB,

$$\begin{bmatrix} \frac{1}{\sqrt{t}} j_t^1 \\ P_3' \end{bmatrix} N N' \begin{pmatrix} \frac{1}{\sqrt{t}} j_t^1, P_3 \end{pmatrix} = \begin{bmatrix} rk & 0 \\ 0 & (r-\lambda)I_{t-1} \end{bmatrix}.$$

Let P_2' be a set of $b-1$ orthogonal vectors such that

$$P_2' N' N P_2 = D_2$$

where D_2 is diagonal with the characteristic roots of $N'N$ other than rk on the diagonal.

Since the non-zero characteristic roots of NN' and $N'N$ are equal and of the same multiplicity, partition P_2 into (P_{21}, P_{22}) such that

$$P_2' N' N P_2 = \begin{bmatrix} P_{21}' \\ P_{22}' \end{bmatrix} N' N (P_{21}, P_{22}) = \begin{bmatrix} 0_{b-t} & 0 \\ 0 & (r-\lambda)I_{t-1} \end{bmatrix}$$

where 0_q denotes a $q \times q$ matrix of zeros. It can be shown that P_{22}' can be expressed in terms of P_3' by the relation

$$P_{22}' = (r-\lambda)^{-1/2} P_3' N$$

Let $A' = (X'_2 - k^{-1}NX'_1)$ and consider $A'A$. The $t-1$ orthogonal vectors, P'_3 also diagonalize $A'A$ (since $A'A = rI - \frac{1}{k} NN'$). Hence

$$P'_3 A' A P_3 = \frac{\lambda t}{k} I_{t-1}.$$

Now let $X = (X_0, X_1, X_2)$. Since XX' is symmetric, there exists an orthogonal matrix Q such that $Q'XX'Q = D$ where D is diagonal. The rank of XX' for the balanced incomplete block design is $b+t-1$. Therefore, if we partition Q into (Q_1, P_4) , where Q_1 and P_4 are of dimension $b+t-1 \times n$ and $(n-b-t+1) \times n$ respectively, we may write

$$\begin{aligned} \begin{bmatrix} Q_1 \\ P_4 \end{bmatrix} XX' (Q_1, P_4) &= \begin{bmatrix} Q_1 \\ P_4 \end{bmatrix} (X_0 X'_0 + X_1 X'_1 + X_2 X'_2) (Q_1, P_4) \\ &= \begin{bmatrix} D_4 & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

Hence
$$P'_4 (X_0 X'_0 + X_1 X'_1 + X_2 X'_2) P_4 = 0.$$

Since the matrices $X_0 X'_0$, $X_1 X'_1$ and $X_2 X'_2$ are positive semi-definite, it follows that P_4 is a set of $n-b-t+1$ orthogonal vectors such that $P'_4 X_0 = 0$, $P'_4 X_1 = 0$ and $P'_4 X_2 = 0$.

We now define an orthogonal matrix P as follows:

$$P' = \begin{bmatrix} n^{-1/2} j_n^1 \\ k^{-1/2} P'_{21} X'_1 \\ k^{-1/2} P'_{22} X'_1 \\ \left(\frac{k}{\lambda t}\right)^{1/2} P'_3 A' \\ P'_4 \end{bmatrix}$$

Consider now

$$g(Y; \theta) = \frac{1}{(2\pi)^{n/2} |V|^{1/2}} \exp -\frac{1}{2} (Y - \bar{\mu})' P P' V^{-1} P P' (Y - \bar{\mu}).$$

By first finding the form of $P'VP$ and using the fact that

$$(P'V^{-1}P) = (P'VP)^{-1}$$

MINIMAL SUFFICIENT STATISTICS FOR BIBD UNDER EISENHART MODEL II
we have the result as shown in Table 1.

TABLE 1. FORM OF $P'V^{-1}P$

$(\sigma^2 + k\sigma_1^2 + r\sigma_2^2)^{-1}$	0	0	0	0
0	$(\sigma^2 + k\sigma_1^2)^{-1}\mathbf{I}_{b-t}$	0	0	0
0	0	$d_1^{-1}(\sigma^2 + k^{-1}\lambda t\sigma_2^2)\mathbf{I}_{t-1}$	$-[k^{-2}\lambda t(r-\lambda)]\frac{1}{2}$	0
0	0	$-[k^{-2}\lambda t(r-\lambda)]\frac{1}{2}d_1^{-1}\sigma_2^2\mathbf{I}_{t-1}$	$[\sigma^2 + k\sigma_1^2 + k^{-1}(r-\lambda)\sigma_2^2]d_1^{-1}\mathbf{I}_{t-1}$	0
0	0	0	0	$\sigma^{-2}\mathbf{I}_{n-b-t+1}$

$$d_1 = \sigma^4 + k\sigma^2\sigma_1^2 + r\sigma^2\sigma_2^2 + \lambda t\sigma_1^2\sigma_2^2$$

The vector $P'(Y - \bar{\mu})$ can be shown to be equal to the following :

$$P'(Y - \bar{\mu}) = \begin{bmatrix} n^{\frac{1}{2}}(y \dots - \mu) \\ k^{-\frac{1}{2}}P'_{21}X_1Y \\ k^{-\frac{1}{2}}P'_{22}X_2Y \\ \left(\frac{k}{\lambda t}\right)^{\frac{1}{2}}P'_3A'Y \\ P'_4Y \end{bmatrix}$$

where $y \dots = \frac{1}{n}j'_nY$.

If we let $q = (Y - \bar{\mu})'PP'V^{-1}PP'(Y - \bar{\mu})$,

we have

$$\begin{aligned} q = & (\sigma^2 + k\sigma_1^2 + r\sigma_2^2)^{-1} n(y \dots - \mu)^2 + [k(\sigma^2 + k\sigma_1^2)]^{-1} Y'X_1P_{21}P_{21}'X_1Y + \\ & + k(\sigma^2 + k^{-1}\lambda t\sigma_2^2)d_1^{-1} Y'X_1P_{22}P'_{22}X_1Y + \sigma^{-2}Y'P_4P_4'Y + \left(\frac{k}{\lambda t}\right) [\sigma^2 + k\sigma_1^2 + \\ & + k^{-1}(r-\lambda)\sigma_2^2]d_1^{-1} Y'AP_3P'_3A'Y - 2[k^{-2}(r-\lambda)\lambda t]d_1^{-1} Y'X_1P_{22}P'_3A'Y. \end{aligned}$$

From the form of q we may define six statistics as follows:

$$\begin{aligned} s_1 &= y \dots \\ s_2 &= k^{-1} \mathbf{Y}' \mathbf{X}_1 \mathbf{P}_{21} \mathbf{P}_{21}' \mathbf{X}_1' \mathbf{Y} \text{ (not defined if } b = t) \\ s_3 &= k^{-1} \mathbf{Y}' \mathbf{X}_1 \mathbf{P}_{22} \mathbf{P}_{22}' \mathbf{X}_1' \mathbf{Y} \\ s_4 &= k^{-1} (r - \lambda) \mathbf{Y}' \mathbf{X}_1 \mathbf{P}_{22} \mathbf{P}_{32}' \mathbf{A}' \mathbf{Y} \\ s_5 &= \left(\frac{k}{\lambda t} \right) \mathbf{Y}' \mathbf{A} \mathbf{P}_3 \mathbf{P}_3' \mathbf{A}' \mathbf{Y} \\ s_6 &= \mathbf{Y}' \mathbf{P}_4 \mathbf{P}_4' \mathbf{Y} \end{aligned} \quad \dots \quad (3.1)$$

These six statistics are, by definition, sufficient for the parameters μ , σ^2 , σ_1^2 , and σ_2^2 .

We shall now prove that this set of sufficient statistics is minimal for $g(\mathbf{Y}; \theta)$.

$g(\mathbf{Y}; \theta)$ may be written in the form

$$g(\mathbf{Y}; \theta) = P(\theta) Q(\mathbf{Y}) e^{-\frac{1}{2} \sum_{i=1}^k v_i(\theta) u_i(\mathbf{Y})}.$$

A necessary and sufficient condition for the set of sufficient statistics $u_i(\mathbf{Y})$ to be minimal for $g(\mathbf{Y}; \theta)$ is that there exist no non-zero constants $a_1, a_2, \dots, a_k, c$ such that

$$\sum_{i=1}^k a_i v_i(\theta) = c. \quad \dots \quad (3.2)$$

Thus it is enough to prove that for the following seven functions :

$$\begin{aligned} v_1 &= (\sigma^2 + k\sigma_1^2 + r\sigma_2^2)^{-1} \\ v_2 &= (\sigma^2 + k\sigma_1^2)^{-1} \\ v_3 &= [\sigma^2 + k\sigma_1^2 + k^{-1}(r - \lambda)\sigma_2^2] d_1^{-1} \\ v_4 &= [\sigma^2 + k^{-1}\lambda t\sigma_2^2] d_1^{-1} \\ v_5 &= -2\sigma_2^2 d_1^{-1} \\ v_6 &= \sigma^{-2} \\ v_7 &= v_1 \mu. \end{aligned} \quad \dots \quad (3.3)$$

(3.2) is not true for any $a_1, a_2, \dots, a_7$ and c except when all of them vanish (Rao, 1945).

In (3.3) it is clear that μ appears only in v_7 . Since $v_1, v_2, \dots, v_6$ are homogeneous functions of σ, σ_1 and σ_2 of degree -2 , the constant c can only be zero.

MINIMAL SUFFICIENT STATISTICS FOR BIBD UNDER EISENHART MODEL II

Effecting the linear transformation $x = \sigma^2$; $y = \sigma^2 + k\sigma_1^2$; $z = \sigma^2 + k\sigma_1^2 + r\sigma_2^2$ the functions in (3.3) become

$$v_1 = xy \left[xz + \frac{\lambda t}{rk} (y-x)(z-y) \right] D^{-1}$$

$$v_2 = xz \left[xz + \frac{\lambda t}{rk} (y-x)(z-y) \right] D^{-1}$$

$$v_3 = xyz \left[y + \frac{(r-\lambda)}{rk} (z-y) \right] D^{-1}$$

$$v_4 = xyz \left[x + \frac{\lambda t}{rk} (z-y) \right] D^{-1}$$

$$v_5 = -2xyz \frac{(z-y)}{r} D^{-1}$$

$$v_6 = yz \left[xz + \frac{\lambda t}{rk} (y-x)(z-y) \right] D^{-1}$$

where
$$D = xyz \left[xz + \frac{\lambda t}{rk} (y-x)(z-y) \right].$$

Observe that the term x^2y^2 appears only in v_1 , x^2z^2 appears only in v_2 , and y^2z^2 appears only in v_6 . This being the case, v_1 , v_2 and v_6 are mutually linearly independent and linearly independent of v_3 , v_4 , and v_5 . Now observe that after removing the common factor xyz in v_3 , v_4 , and v_5 , these are also linearly independent thereby proving that (3.2) is not true unless $a_1, a_2, \dots, a_6$ and c vanish. This condition then implies that the set of sufficient statistics defined in (3.1) is minimal.

In Table 2, the relationship of these six statistics to the intra-block analysis of variance (Kempthorne, 1952), the expectation of the statistics, the pairwise independence, and the distribution of each is given. Concerning Table 2,

(a) the relationship for s_2 holds only if $b > t$. s_2 is not defined if $b = t$ and in which case there are but five statistics in a minimal set of sufficient statistics;

(b) $\chi^2(p)$ denotes the central chi-square distribution with p degrees of freedom;

(c) the ρ_i are the non-zero characteristic roots of $\frac{1}{2}(A_4 + A_4')$ where $A_4 = k^{-1}X_1N'(X_2' - k^{-1}NX_1')$;

(d) B_i is the i -th block total, T_j is the total of all blocks containing treatment j , V_j is the j -th treatment total and $Q_j = (V_j - k^{-1}T_j)$. (Note also that Q_j has no relationship to the orthogonal matrix Q_1 mentioned previously).

Then we can define a random variable X taking non-negative integral values in T with probabilities

$$P_x = \text{Prob}\{X = x\} = \frac{a_x \theta^x}{f(\theta)} \quad x \in T \quad \dots (1.3)$$

and call this distribution analogously a generalized power series distribution (gpsd). It may be noted that gpsd reduces to a psd when T is the entire set of non-negative integers. The properties established by Noack (1950) and Khatri (1959) for psd can be easily deduced for gpsd by following the same lines. Further, it can be easily seen that proper choice of T and $f(\theta)$ reduces the gpsd, in particular, to the Binomial, negative Binomial, Poisson and logarithmic series distributions and their truncated forms. Incidentally, it is obvious that truncated gpsd is itself a gpsd in its own right and hence the properties that hold for a gpsd continue to hold for its truncated forms.

Problems of statistical inference associated with psd's do not seem to have been much investigated. Roy and Mitra (1957) have derived the uniformly minimum variance unbiased estimates in certain particular cases and have provided necessary tables for Poisson distribution truncated at zero. The author (1957) has shown that for gpsd (1.2), the maximum likelihood method and the method of moments give the same estimate of the parameter of the gpsd. The likelihood equation and a method for solving it are derived for the problem of estimation. In this paper, we suggest what we call the, "Ratio Method" for estimation of the parameter of the gpsd and investigate its important properties and study certain applications.

2. ESTIMATION BY THE RATIO METHOD FOR A G.P.S.D.

[Range T finite and $T = (c, c+1, \dots, c+k = d)$ with positive probabilities].

The gpsd that we consider here is of the form :

$$P_x = \text{Prob}\{X = x\} = \frac{a_x \theta^x}{f(\theta)} \quad \dots (2.1)$$

where

$$x \in T = (c, c+1, \dots, c+k = d), \quad d \text{ finite}$$

$$f(\theta) = \sum_{x=c}^d a_x \theta^x \quad \dots (2.2)$$

and

$$a_x > 0 \text{ for } x \in T.$$

Let

$$g_r(x) = \frac{a_{x-r}}{a_x} \quad x \in T \quad \dots (2.3)$$

with r being an integer such that $x-r \in T$. Then

$$\sum_{x=u}^v g_r(x) P_x = \theta^r \sum_{x=u-r}^{v-r} P_x \quad \dots (2.4)$$

ON GENERALIZED POWER SERIES DISTRIBUTIONS

where u and v are arbitrary with $c+r \leq u \leq v \leq d$. From (2.4) we get the identity

$$\theta^r = \sum_{x=u}^v g_r(x) P_x \bigg/ \sum_{x=u-r}^{v-r} P_x \quad \dots \quad (2.5)$$

which can be made use of in problems of estimation. In a sample of size N , if n_x is the observed frequency for x , then since $E(n_x) = NP_x$, the statistic

$$\sum_{x=u}^v g_r(x) n_x \bigg/ \sum_{x=u-r}^{v-r} n_x \quad \dots \quad (2.6)$$

may be taken as an estimate of θ^r for admissible values of $r = 1, 2$, etc. Since u and v are arbitrary, the same method is applicable for estimation in truncated and censored gpsd's also, provided that their range contains a subset of consecutive integers. We call these estimates "ratio estimates."

It is interesting to note that Plackett (1953) and Moore (1952, 1954) applied this ratio method to the special cases of estimating θ in truncated Binomial and Poisson distributions. The method which we call the ratio method is applicable not merely for estimating θ , but also for its integral powers and for any gpsd of this section, truncated or censored.

The ratio estimate is not generally unbiased or efficient, but is always easy to compute. In certain cases (see Section 3), however, unbiased estimates can be obtained by the ratio method. In other cases, such as those in this section, the bias is generally of the order $\frac{1}{N}$. It may be easily verified that no unbiased estimate for θ (and θ^r in general) exists when the range of T is finite as in situations considered here.

Consider the following ratio estimate of θ for gpsd (2.1):

$$\theta' = \frac{t_1}{t_2} \quad \dots \quad (2.7)$$

where

$$t_1 = \sum_{x=c+1}^d \left(\frac{a_{x-1}}{a_x} \right) n_x \quad \dots \quad (2.8)$$

and

$$t_2 = \sum_{x=c}^{d-1} n_x \quad \dots \quad (2.9)$$

Then, writing

$$E(t_2) = N \sum_{x=c}^{d-1} P_x = N(1 - P_d) = NP, \text{ say,} \quad \dots \quad (2.10)$$

where

$$P = 1 - P_d, \quad \dots \quad (2.11)$$

we have

$$E(t_1) = NP\theta. \quad \dots \quad (2.12)$$

Let

$$t_1 - E(t_1) = \delta t_1 \quad \text{and} \quad t_2 - E(t_2) = \delta t_2. \quad \dots \quad (2.13)$$

Then

$$\theta' = \frac{t_1}{t_2} = \theta \left(1 + \frac{\delta t_1}{NP\theta} \right) \left(1 + \frac{\delta t_2}{NP} \right)^{-1}.$$

Since the deviations $\delta t_1, \delta t_2$ are stochastically of order $N^{\frac{1}{2}}$, we get on expansion

$$\theta' = \theta \left[1 + \frac{\delta t_1}{NP\theta} - \frac{\delta t_2}{NP} - \frac{(\delta t_1)(\delta t_2)}{N^2 P^2 \theta} + \frac{(\delta t_2)^2}{N^2 P^2} \right] \quad \dots (2.14)$$

neglecting terms of order higher than $\frac{1}{N}$. Thus, to this order of approximation,

$$E(\theta') = \theta \left[1 + \frac{E(\delta t_2)^2}{N^2 P^2} - \frac{E(\delta t_1)(\delta t_2)}{N^2 P^2 \theta} \right]. \quad \dots (2.15)$$

Now a little computation gives

$$E(\delta t_2)^2 = NP(1-P) \quad \dots (2.16)$$

and

$$E(\delta t_1)(\delta t_2) = N\theta[P(1-P) - P_{d-1}]. \quad \dots (2.17)$$

Thus

$$E(\theta') = \theta + \frac{1}{N} \left(\frac{\theta P_{d-1}}{P^2} \right) \quad \dots (2.18)$$

from which we get the magnitude of the bias θ' , to order $\frac{1}{N}$,

$$b(\theta') = \frac{1}{N} \left(\frac{\theta P_{d-1}}{P^2} \right) = (\theta P_{d-1})/N(1-P_d)^2 = \frac{B(\theta')}{N}, \text{ say.} \quad \dots (2.19)$$

The variance of θ' correct to terms of order $\frac{1}{N}$ is

$$\text{Var}(\theta') = \frac{1}{N^2 P^2} [E(\delta t_1)^2 + \theta^2 E(\delta t_2)^2 - 2\theta E(\delta t_1)(\delta t_2)]. \quad \dots (2.20)$$

Now

$$E(\delta t_1)^2 = N(D - P^2 \theta^2) \quad \dots (2.21)$$

where

$$D = \sum_{x=c+1}^d \left(\frac{a_{x-1}}{a_x} \right)^2 P_x. \quad \dots (2.22)$$

Thus, to order $\frac{1}{N}$

$$\text{Var}(\theta') = \frac{1}{NP^2} [D - P\theta^2 + 2\theta^2 P_{d-1}]. \quad \dots (2.23)$$

3. UNBIASED ESTIMATION BY THE RATIO METHOD FOR A G.P.S.D.

[Range T infinite and $T = (c, c+1, \dots)$ with positive probabilities].

Uniformly minimum variance unbiased estimation for psd's has been considered by Roy and Mitra (1957). Tate and Goen (1958) have considered the same for truncated Poisson.

ON GENERALIZED POWER SERIES DISTRIBUTIONS

It is easy to demonstrate that the ratio method discussed in Section 2 gives the unique unbiased estimate of θ , linear in frequencies, for a gpsd with range T infinite and $T = (c, c+1, \dots)$ with positive probabilities. For, consider the gpsd

$$P_x = \text{Prob} \{X = x\} = \frac{a_x \theta^x}{f(\theta)} \quad x = c, c+1, \dots \quad \dots \quad (3.1)$$

where
$$f(\theta) = \sum_{x=c}^{\infty} a_x \theta^x \quad \dots \quad (3.2)$$

and
$$a_x > 0 \quad \text{for all } x = c, c+1, \dots$$

Now, if in a sample of size N from gpsd (3.1), the frequency of x is n_x and we want an unbiased estimate for θ of the type linear in n_x , we should be able to demonstrate the existence of a function of x , $t(x)$, such that, denoting the corresponding estimate

$$\tilde{\theta} = \sum_{x=c}^{\infty} t(x) n_x \quad \dots \quad (3.3)$$

we must have $E(\tilde{\theta}) = \theta$ for all θ in the parameter space of (3.1). That is

$$N \sum_{x=c}^{\infty} t(x) a_x \theta^x = \sum_{x=c}^{\infty} a_x \theta^{x+1}.$$

Since this is an identity in θ , equating coefficients of corresponding powers of θ , we get

$$t(x) = \begin{cases} 0 & \text{for } x = c \\ \frac{1}{N} \left(\frac{a_{x-1}}{a_x} \right) & \text{for } x = c+1, c+2, \dots \end{cases}$$

Thus, the unique unbiased estimate of θ linear in the frequencies comes out to be the ratio estimate θ' . The exact variance of this estimate is

$$\sigma^2(\theta') = \frac{1}{N} \left[\sum_{x=c+1}^{\infty} \left(\frac{a_{x-1}}{a_x} \right)^2 P_x - \theta^2 \right]. \quad \dots \quad (3.4)$$

An unbiased estimate of $\sigma^2(\theta')$ is

$$\left[\sum_{x=c+1}^{\infty} \left(\frac{a_{x-1}}{a_x} \right)^2 n_x - N \theta'^2 \right] / N(N-1) \quad \dots \quad (3.5)$$

the proof of which is almost immediate once one recognizes that θ' is the mean of N independent identically distributed random variables Y_i with probability distribution given by (for $i = 1, 2, \dots, N$)

$$\text{Prob } \{Y_i = 0\} = P_c$$

and
$$\text{Prob } \left[Y_i = \frac{a_{x-1}}{a_x} \right] = P_x \text{ for } x = c+1, c+2, \dots$$

One can compare $\sigma^2(\theta')$ with the asymptotic variance $\text{Var}(\hat{\theta})$ of the maximum likelihood estimate of θ and the efficiency of the ratio estimate θ' can be computed.

Lastly, one may establish that

$$\hat{\theta}' = \frac{1}{N} \sum_{x=c+r}^{\infty} \left(\frac{a_{x-r}}{a_x} \right) n_x \quad \dots (3.6)$$

is the only unbiased estimate of θ' (r an integer) which is a linear function of the frequencies.

4. ESTIMATION BY THE RATIO METHOD FOR SINGLY TRUNCATED BINOMIAL DISTRIBUTION

Fisher (1936) and Haldane (1932, 1938) discussed uses of the truncated binomial distribution. For instance, in problems of human genetics, in estimating the proportion of albino children produced by couples capable of producing albinos, sampling has necessarily to be restricted to families having at least one albino child. Finney (1949) has cited some more applications. Fisher and Haldane derived the maximum likelihood procedure to estimate the parameter π . Patil (1959) gave a direct method to obtain the maximum likelihood estimate. Moore (1954) suggested a simple "ratio-estimate" based on an identity between binomial probabilities. For a slightly different problem, where, in a sample from a complete binomial distribution, the frequencies in some lowest classes are missing, Rider (1955) suggested a method of estimation, which uses first two moments of the complete binomial and leads to a linear equation.

The probability law of the binomial distribution truncated at c on the left can be written as

$$b^*(x, \pi, n) = \left(B^*(c, \pi, n) \right)^{-1} \binom{n}{x} \pi^x (1-\pi)^{n-x}, \quad x = c, c+1, \dots, n. \quad \dots (4.1)$$

where

$$B^*(c, \pi, n) = \sum_{x=c}^n \binom{n}{x} \pi^x (1-\pi)^{n-x} \quad \dots (4.2)$$

The first two moments about the origin of (4.1), then, are

$$\mu^* = \mu^*(c, \pi, n) = n\pi. \quad B^*(c-1, \pi, N-1)/B^*(c, \pi, n) \quad \dots (4.3)$$

and

$$m_2^* = m_2^*(c, \pi, n) = \mu^*(c, \pi, n) \{1 + \mu^*(c-1, \pi, n-1)\}. \quad \dots (4.4)$$

The case of truncation to the right can be dealt with in a similar way by replacing π by $1-\pi$ and the truncation point c by $n-c$.

ON GENERALIZED POWER SERIES DISTRIBUTIONS

In this case $a_{x-1} | a_x = x | (n-x+1)$ and since $\theta = \pi/(1-\pi)$, (see Patil, 1959), we have the following "ratio-estimate" for π :

$$\pi' = \frac{t_1}{t_1 + t_2} \quad \dots (4.5)$$

where
$$t_1 = \sum_{x=c+1}^n \left(\frac{x n_x}{n-x+1} \right) \text{ and } t_2 = \sum_{x=c}^{n-1} n_x.$$

To investigate the efficiency of π' given by (4.5) its asymptotic variance can be written down as:

$$\text{Var} (\pi') = \frac{(1-\pi)^2}{NP^2} [(1-\pi)^2 D - P\pi^2 + 2\pi^2 \cdot P_{n-1}] \quad \dots (4.6)$$

where
$$P = \sum_{n=c}^{n-1} b^*(x, \pi, n)$$

$$D = \sum_{x=c+1}^n \left(\frac{x}{n-x+1} \right) b^*(x, \pi, n)$$

and
$$P_{n-1} = b^*(n-1, \pi, n) .$$

Also the asymptotic variance of the maximum likelihood estimate $\hat{\pi}$, (Patil, 1959), is given by

$$\text{Var} (\hat{\pi}) = \frac{\pi(1-\pi)^2}{N\mu_2^*}$$

where μ_2^* is the variance of (4.1).

Therefore the asymptotic efficiency of π' takes the form :

$$\text{Eff} (\pi') = \frac{P^2}{\mu_2^*} \left(\left(\frac{1-\pi}{\pi} \right)^2 D - P + 2P_{n-1} \right)^{-1} . \quad \dots (4.7)$$

The special cases of some importance in genetics are $c = 1$ and $\pi = 1/4, 1/2$, or $3/4$. The efficiency of the Ratio-Estimate (R) relative to the Maximum Likelihood Estimate (ML) in these cases is tabulated and shown in Table 1.

TABLE 1. ASYMPTOTIC EFFICIENCY OF R FOR $c = 1$

n	$\pi = 1/4$	$1/2$	$3/4$
3	.924	.875	.875
4	.909	.769	.772
5	.919	.715	.664
6	.933	.694	.563
7	.947	.693	.523
8	.952	.705	.481
9	.956	.723	.435
10	.959	.776	.388

Examination of the above table shows that the efficiency of R in case of $\pi = 1/4$ and $\pi = 1/2$ decreases in the beginning with n , reaches a minimum and then increases with increasing values of n . For $\pi = 3/4$, however, the efficiency decreases throughout for $n = 3(1)10$.

Following Section 2, one gets, to order $1/N$, the amount of bias of π' (R) as follows:

$$b(\pi') = \frac{(1-\pi)^2}{NP^2} [\pi^2 P + (\pi - 2\pi^2)P_{n-1} - (1-\pi)^2 D] = \frac{B(\pi')}{N}.$$

The table below gives $B(\pi')$ for $c = 1$ and $\pi = 1/4, 1/2$ and $3/4$.

TABLE 2. N (AMOUNT OF BIAS TO ORDER $1/N$) OF R

n	$\pi = 1/4$	$1/2$	$3/4$
3	-.1927	.1458	.0820
4	.1227	.1307	.0833
5	.0861	.1184	.0893
6	.0648	.1062	.0962
7	.0510	.0943	.0931
8	.0420	.0833	.0928
9	.0355	.0736	.0915
10	.0301	.0651	.0896

Table 2 shows that R is an underestimate of π . A closer investigation, however, brings out that the bias to order $1/N$ is quite small for R. One may note that the maximum likelihood estimate also happens to be biased in this case (Patil, 1959).

Illustrative example. The detailed computation procedure of the ratio estimate discussed above will be illustrated with reference to K. Pearson's data on albinism in man. The table below gives the number of families (n_x), each of five children having exactly x children in the family, ($x = 1, 2, 3, 4, 5$).

number of albinos in family (x)	1	2	3	4	5
number of families (n_x)	25	23	10	1	1

If π is the probability for a child to be an albino, we may accept the truncated binomial model:

$$\frac{\binom{n}{x} \pi^x (1-\pi)^{n-x}}{1 - (1-\pi)^n} \quad x = 1, 2, \dots, n$$

ON GENERALIZED POWER SERIES DISTRIBUTIONS

for the probability of x albinos in a family of n . Here $n = 5$, and the problem is to estimate π on the basis of the data given in the table above.

Here, $t_2 = \sum_{x=1}^{n-1} n_x = 59$ and t_1 can be computed from the following table :

x	$\frac{x}{n-x+1}$	n_x
2	.5	23
3	1	10
4	2	1
5	5	1

Then,
$$t_1 = \sum_{x=2}^n \left(\frac{x}{n-x+1} \right) n_x = 28.50.$$

The ratio estimate is obtained as $\pi' = \frac{28.50}{28.50+59} = 0.3257$.

To compute the variance of π' , we require

$$P = 1 - \frac{\pi^n}{1-(1-(1-\pi))^n} = 0.99574$$

$$P_{n-1} = \frac{n(1-\pi)}{\pi} (1-P) = 0.04408$$

and

$$D = \left(\frac{\pi}{1-\pi} \right) \left[\frac{n+1}{1-(1-\pi)^n} \left\{ (1-\pi^n) E \left(\frac{1}{x}, n, 1-\pi \right) - \frac{(1-\pi)^n}{n} \right\} - P \right] = 0.56156$$

where
$$E \left(\frac{1}{x}, n, \pi \right) = \sum_{x=1}^n \frac{1}{x} \binom{n}{x} \pi^x (1-\pi)^{n-x} / \{1-(1-\pi)^n\}$$

and is tabulated by Grab and Savage (1954).

By linear interpolation from the table by Grab and Savage, $E \left(\frac{1}{x}, n, 1-\pi \right) = 0.33870$ taking $\pi' = 0.3257$ as the estimate for π throughout.

Then the variance of π' is estimated from the formula

$$\text{Var}(\pi') = \frac{(1-\pi)^2}{NP^2} [(1-\pi)^2 D - P\pi^2 + 2\pi^2 \cdot P_{n-1}] = 0.0013410$$

so that the standard error of π' is S.E. (π') = 0.03662.

Incidentally, the maximum likelihood estimate $\hat{\pi}$ comes out (Patil, 1959) to be in this case, $\hat{\pi} = 0.3088$ with S.E. ($\hat{\pi}$) = 0.03210.

6. ESTIMATION BY RATIO METHOD FOR TRUNCATED POISSON DISTRIBUTION

Problems of estimation in a truncated Poisson distribution with known truncation points have been discussed by various authors. The case of truncation on the left has been considered by David and Johnson (1948) who gave the maximum likelihood estimate, by Plackett (1953) who gave a simple and highly efficient ratio estimate, and by Rider (1953) who used first two moments. Truncation on the right has been discussed by Tippet (1932), Bliss (1948), and Moore (1952). Tippet derived the maximum likelihood solution; Bliss developed an approximation to it; and Moore suggested a simple ratio estimate. Uniformly minimum variance unbiased estimates have been obtained by Roy and Mitra (1957) and by Tate and Goen (1958). For both types of truncations, the author (1959) has provided neat and compact equations for estimation by the method of maximum likelihood. He has also presented numerical tables and some suitable charts to facilitate the solution of these equations in certain special cases. In this section, we study the Ratio Method as applied to truncated Poisson distributions.

The probability law of the singly truncated Poisson distribution with truncation point on the right at d can be written as:

$$p^*(x, \mu) = [P(d, \mu)]^{-1} e^{-\mu} \frac{\mu^x}{x!} \quad x = 0, 1, 2, \dots, d \quad \dots (6.1)$$

where

$$P(d, \mu) = \sum_{x=0}^d e^{-\mu} \frac{\mu^x}{x!} \quad \dots (6.2)$$

In this case, $a_{x-1}/a_x = x$, and since $\theta = \mu$, the ratio estimate for μ takes the form

$$\mu' = \frac{\sum_{x=0}^d x n_x}{\sum_{x=0}^{d-1} n_x} \quad \dots (6.3)$$

as first suggested by Moore (1954). The following table gives the asymptotic efficiency of μ' relative to $\hat{\mu}$ for values of $d = 5$ with $\mu = .25, .5(.5)2.5$, and $d = 10$ with $\mu = .5(.5)4.5$.

TABLE 3. EFFICIENCY OF R

μ	.25	.50	1.00	1.50	2.00	2.50
Case (i) $d = 5$						
Eff.	.999	.990	.979	.967	.951	.923

μ	.5	1.0	1.5	2.0	2.5	3.0	3.5	4	4.5
Case (ii) $d = 10$									
Eff.	1.000	1.000	1.000	1.000	.999	.992	.981	.894	.817

ON GENERALIZED POWER SERIES DISTRIBUTIONS

Thus, R seems to be highly efficient on the whole and its efficiency always exceeds 82 percent in situations considered in Table 3.

The following table gives $B(\mu')$ for values of $d = 5$ with $\mu = .25, .5(.5)2.5$ and $d = 10$ with $\mu = .5(.5)5$.

TABLE 4. N (AMOUNT OF BIAS TO ORDER $1/N$) OF R

μ	.25	.50	1.00	1.50	2.00	2.50				
Case (i) $d = 5$										
$B(\mu')$	.0003	.0008	.0015	.0719	.1977	.4461				
Case (ii) $d = 10$										
μ	.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$B(\mu')$	.0000	.0000	.0003	.0004	.0022	.0081	.0231	.0536	.1063	.1876

Table 4 shows that, though over-estimate, R involves almost negligible bias.

The probability law of the singly truncated Poisson distribution with truncation point on the left at c can be written as :

$$p^*(x, \mu) = [P^*(c, \mu)]^{-1} e^{-\mu} \frac{\mu^x}{x!} \quad x = c, c+1, \dots \infty \quad \dots (6.8)$$

where

$$P^*(c, \mu) = \sum_{x=c}^{\infty} e^{-\mu} \frac{\mu^x}{x!} \quad \dots (6.9)$$

In this case, $a_{x-1}/a_x = x$ and since $\theta = \mu$, we have the following "ratio estimate" for μ :

$$\mu' = \frac{\sum_{x=c+1}^{\infty} x n_x}{N} \quad \dots (6.10)$$

when $c = 1$, i.e., when only "zero" counts are truncated, the estimate takes the form suggested by Plackett (1953):

$$\mu' = \frac{\sum_{x=2}^{\infty} x n_x}{N}. \quad \dots (6.11)$$

The unique unbiased estimate of μ linear in the frequencies (*ibid.*, Section 3) is provided in (6.10). The exact variance of this estimate is

$$\sigma^2(\mu') = \frac{1}{N} \left[\sum_{x=c+1}^{\infty} x^2 P_x - \mu^2 \right] \quad \dots (6.12)$$

and an unbiased estimate of $\sigma^2(\mu')$ is

$$\left\{ \sum_{x=c+1}^{\infty} x^2 n_x - N \mu'^2 \right\} / N(N-1) \quad \dots (6.13)$$

when $c = 1$, (6.12) reduces to

$$\sigma^2(\mu') = \frac{1}{N} [\mu + \mu^2/(e^\mu - 1)] \quad \dots (6.14)$$

first derived by Plackett (1953). Plackett computed also the efficiency of μ' in this special case. The following table gives the efficiencies of μ' relative to $\hat{\mu}$ as computed by Plackett.

TABLE 5. EFFICIENCY OF R FOR $c = 1$.

μ	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
Eff.	.9693	.9559	.9539	.9586	.9662	.9743	.9815	.9872

The author is thankful to Professor C. R. Rao and Dr. J. Roy for helpful discussions at the Indian Statistical Institute and to the referee for helpful comments.

REFERENCES

- DAVID, F. N. and JOHNSON, N. L. (1952): Truncated Poisson. *Biometrics*, **8**, 275-285.
- FISHER, R. A. (1936): The effects of method of ascertainment upon estimation of frequencies. *Ann. Eugen.*, **6**, 13-25.
- FINNEY, D. J. (1949): The truncated binomial distribution. *Biometrics*, **9**, 176-200.
- FINNEY, D. J. and VARLEY, G. C. (1955): An example of the truncated Poisson distribution. *Biometrics*, **10**, 417-451.
- GRAB, E. L. and SAVAGE, I. R. (1954): Tables of the expected value of $1/X$ for positive Bernoulli and Poisson variables. *J. Amer. Stat. Ass.*, **49**, 169-177.
- HALDANE, J. B. S. (1941): The fitting of binomial distribution. *Ann. Eugen.*, **11**, 179-181.
- (1945): On a method of estimating frequencies. *Biometrika*, **33**, 222-225.
- MOORE, P. G. (1952): The estimation of the Poisson parameter from a truncated distribution. *Biometrika*, **39**, 247-251.
- (1954): A note on truncated Poisson distributions. *Ann. Math. Stat.*, **14**, 389-400.
- PATIL, G. P. (1957): Thesis submitted for the Associateship of the Indian Statistical Institute.
- (1959): Doctoral Thesis submitted to the University of Michigan.
- PEARSON, K. (1913): A monograph on albinism in man. *Drapers Company Research Memoirs*.
- PLACKETT, R. L. (1953): The truncated Poisson distribution. *Biometrics*, **9**, 485-488.
- RIDER, P. R. (1953): Truncated Poisson distributions. *J. Amer. Stat. Ass.*, **48**, 826-830.
- (1955): Truncated binomial and negative binomial distributions. *J. Amer. Stat. Ass.*, **50**, 877-883.
- ROY, J. and MITRA, S. K. (1957): Unbiased minimum variance estimation in a class of discrete distributions. *Sankhyā*, **18**, 371-378.
- TATE and GOEN (1958): Minimum variance unbiased estimation for the truncated Poisson distribution. *Ann. Math. Stat.*, **29**, 755-765.
- TIPPETT, L. H. C. (1932): A modified method of counting particles. *Proc. Roy. Soc., Series A*, **137**, 434-446

Paper received : April, 1960.

Revised : August, 1960.

METHOD OF CONSTRUCTING TWO MUTUALLY ORTHOGONAL LATIN SQUARES OF ORDER $3n+1$ *

By P. KESAVA MENON

Ministry of Defence, New Delhi

SUMMARY. Bose, Shrikhande and Parker (1960) have shown that a pair of mutually orthogonal latin squares of all orders other than 2 and 6 can be constructed. They have shown in particular how a pair of orthogonal latin squares of order $3n+1$ where n is odd can be constructed with the help of an orthogonal array.

Given below is a direct method of making two orthogonal squares of order $3n+1$ ($n \neq 2, 6$).

Let A be the circulant matrix

$$A = \begin{pmatrix} 0 & 1 & 2 & \dots & 2n \\ 2n & 0 & 1 & \dots & 2n-1 \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ 1 & 2 & 3 & \dots & 0 \end{pmatrix}$$

Let B be the matrix whose first $n+1$ diagonals (counted from the principal diagonal onwards from left to right) are respectively the first $n+1$ rows of A and whose remaining n diagonals have the constant elements $2n+1, 2n+2, \dots, 3n$ respectively :

$$B = \begin{pmatrix} 0 & 2n & 2n-1 & \dots & n+2 & n+1 & 2n+1 & 2n+2 & \dots & 3n-1 & 3n \\ 3n & 1 & 0 & \dots & n+4 & n+3 & n+2 & 2n+1 & \dots & 3n-2 & 3n-1 \\ 3n-1 & 3n & 2 & \dots & n+6 & n+5 & n+4 & n+3 & \dots & 3n-3 & 3n-2 \\ . & . & . & \dots & . & . & . & . & \dots & . & . \\ . & . & . & \dots & . & . & . & . & \dots & . & . \\ . & . & . & \dots & . & . & . & . & \dots & . & . \\ 2n-5 & 2n-6 & 2n-7 & \dots & 2n+2 & 2n+3 & 2n+4 & 2n+5 & \dots & 2n-3 & 2n-4 \\ 2n-3 & 2n-4 & 2n-5 & \dots & 2n+1 & 2n+2 & 2n+3 & 2n+4 & \dots & 2n-1 & 2n-2 \\ 2n-1 & 2n-2 & 2n-3 & \dots & n & 2n+1 & 2n+2 & 2n+3 & \dots & 3n & 2n \end{pmatrix}$$

**Editorial Note* : It may be seen that the author arrives at the same type of orthogonal squares as those constructed by Bose, Shrikhande and Parker. However, the author provides an elegant demonstration of the existence of these squares.

Let C be the $(2n+1) \times n$ -matrix whose columns are the last n rows of A reckoned from the bottom upwards :

$$C = \begin{pmatrix} 1 & 2 & \dots & n \\ 2 & 3 & \dots & n+1 \\ 3 & 4 & \dots & n+2 \\ \vdots & \vdots & \ddots & \vdots \\ 2n & 0 & \dots & n-2 \\ 0 & 1 & \dots & n-1 \end{pmatrix}$$

Let D be the $n \times (2n+1)$ matrix formed by the row of A beginning with 2, 4, 6, ..., $2n$ respectively:

$$D = \begin{pmatrix} 2 & 3 & 4 & \dots & 2n & 0 & 1 \\ 4 & 5 & 6 & \dots & 1 & 2 & 3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 2n & 0 & 1 & \dots & 2n-3 & 2n-1 & 2n+1 \end{pmatrix}$$

Finally, let L_1, L_2 be two mutually orthogonal latin squares of order n formed by $2n+1, 2n+2, \dots, 3n$.

Then the matrices

$$M_1 = \begin{pmatrix} B & C \\ D & L_1 \end{pmatrix}, \quad M_2 = \begin{pmatrix} B' & D' \\ C' & L_2 \end{pmatrix},$$

form two mutually orthogonal latin squares of order $3n+1$.

The verification is left to the reader.

REFERENCE

- BOSE, R. C., SHRIKHANDE, S. S. and PARKER, E. T. (1960): Further results on the construction of mutually Orthogonal Latin Squares and the falsity of Euler's conjecture. *Can. J. Math.*, **12**, 189-203.

Paper received : May, 1961.

COMBINATORIAL ARRANGEMENTS ANALOGOUS TO ORTHOGONAL ARRAYS

By C. RADHAKRISHNA RAO

Indian Statistical Institute

SUMMARY. In this paper are considered some combinatorial arrangements analogous to orthogonal arrays introduced by the author several years ago (Rao, 1946a). In an orthogonal array, all combinations of s elements taken d at a time allowing repetitions occur an equal number of times in every set of d rows of the array. By not allowing repetitions and relaxing the condition that the combinations should be ordered certain new arrangements called orthogonal arrays of Type I and Type II have been obtained. Their relationships with orthogonal latin squares and their use in the construction of BIB designs have been discussed.

1. INTRODUCTION

In a recent paper the author (Rao, 1961a) considered the construction of BIB designs with replications 11 to 15. Some solutions depended on certain types of arrangements analogous to orthogonal arrays introduced earlier (Rao, 1946a, p.134), developed and applied in the construction of confounded factorial designs (Rao, 1946b, 1947, 1948). The recent significant work of Bose, Shrikhande and Parker (1960) and Parker (1958) on the falsity of Euler's conjecture involved the concepts of orthogonal arrays and closely related arrangements which may be called orthogonal arrays of Type I. The object of the present paper is to examine certain other arrangements called orthogonal arrays of Type II which have been found useful in the construction of dicyclic solutions to BIB designs (Rao, 1961a).

2. DIFFERENT TYPES OF ORTHOGONAL ARRAYS

Consider a set S of s symbols and a $t \times N$ matrix of elements of S . Such a matrix is called :

(i) *an orthogonal array of strength d , constraints t and index λ , and represented by (N, t, s, d) , if in every set of d rows, the N columns contain each of the s^d ordered combinations of s elements taken d at a time allowing repetitions, λ times;*

(ii) *an orthogonal array of Type I, strength d , constraints t and index λ , and represented by $(N, t, s, d) : \text{I}$, if in every set of d rows, the N columns contain each of the $s!/(s-d)!$ ordered combinations of s elements taken d at a time without repetitions, λ times;*

(iii) *an orthogonal array of Type II, strength d , constraints t and index λ , and represented by $(N, t, s, d) : \text{II}$, if in every set of d rows, the N columns contain each of the $s!/d!(s-d)!$ combinations of s elements taken d at a time with order ignored and without repetitions, λ times.*

For example, when $s = 3$ orthogonal arrays of strength 2 are as follows :

	A	A	A	B	B	B	C	C	C
	A	B	C	A	B	C	A	B	C
$(9, 4, 3, 2)$	A	B	C	B	C	A	C	A	B
	A	B	C	C	A	B	B	C	A

	A	B	C	A	B	C
$(6, 3, 3, 2) : I$	B	C	A	C	A	B
	C	A	B	B	C	A

	A	B	C
$(3, 3, 3, 2) : II$	B	C	A
	C	A	B

For orthogonal arrays of strength 2, the minimum value of N and the maximum value of t , for given s , are

array	min N	max t
orthogonal	s^2	$s+1$
„ type I	$s(s-1)$	s
„ type II	$\frac{s(s-1)}{2}$ for s odd $s(s-1)$ for s even	s

In the present paper it is shown that arrays of Type II and strength 2 can be constructed with the optimum values of N and t when a field with s elements, $GF(s)$ exists. For other values of s , the methods used by Parker (1958), and Bose, Shrikhande and Parker (1960) for the construction of orthogonal arrays may be employed.

3. ARRAYS OF TYPE I AND STRENGTH 2

It is known (Rao, 1946a) that $(s^2, t, s, 2) \implies$ the existence of $(t-2)$ mutually orthogonal latin squares (m.o.l.'s). Theorem 1 contains some results establishing the relationship between Type I arrays and m.o.l.'s.

Theorem 1: (i) $(t-1)$ m.o.l.'s of size $s \implies (s(s-1), t, s, 2) : I$
(ii) $(s(s-1), t, s, 2) : I \implies (t-2)$ m.o.l.'s of size s
(iii) $(s^2, t, s, 2) \implies (s(s-1), t-1, s, 2) : I$
(iv) $(t-2)$ m.o.l.'s of size s
with a common directrix $\iff (s(s-1), t, s, 2) : I.$

COMBINATORIAL ARRANGEMENTS ANALOGOUS TO ORTHOGONAL ARRAYS

A latin square of size s is said to have a directrix if s different symbols can be found in s cells with the property that no two cells are in the same row or column. Result (iv) of Theorem 1 is important and has been used earlier in the construction of Quasi-factorial designs (Rao, 1956) and confounded designs for asymmetrical factorial experiments (Nair and Rao, 1948). In the latter paper it has been shown that $(36, 3, 6, 2) : I$ exists although there are no orthogonal squares of size 6 since a latin square of size 6 with a directrix can be written down. It is interesting to note that the method of construction given by Bose, Shrikhande and Parker (1960) always leads to orthogonal squares with a common directrix.

4. ARRAYS OF TYPE II AND STRENGTH 2

Theorem 2 contains the main result in the construction of orthogonal arrays of Type II.

Theorem 2: *Existence of $GF(s) \implies (s(s-1)/2, s, s, 2) : II$ when s is odd.*

When s is even, the minimum value of N is $s(s-1)$ and a Type I array can be constructed with this value of N if $GF(s)$ exists, which is also a Type II array with index 2. The only interesting case is when s is odd.

Let $\alpha_0 = 0, \alpha_1, \dots, \alpha_{s-1}$ be the elements of $GF(s)$. They can also be written as $\alpha_0 = 0, \beta_1, \dots, \beta_{(s-1)/2}, -\beta_1, \dots, -\beta_{(s-1)/2}$ when s is odd. Consider the vectors

$$(\beta_i \alpha_0, \beta_i \alpha_1, \dots, \beta_i \alpha_{s-1}); \quad i = 1, \dots, (s-1)/2. \quad \dots (4.1)$$

The symmetric differences of elements occurring in the $(r+1)$ -th and $(u+1)$ -th positions are

$$\beta_i(\alpha_r - \alpha_u), -\beta_i(\alpha_r - \alpha_u); \quad i = 1, \dots, (s-1)/2. \quad \dots (4.2)$$

Since $(\alpha_r - \alpha_u) \neq 0$, the differences (4.2) include all the non-zero elements of $GF(s)$ exactly once each. Hence by an application of a theorem due to Bose (1939) it follows that the totality of sets

$$(\beta_i \alpha_0 + \alpha_j, \beta_i \alpha_1 + \alpha_j, \dots, \beta_i \alpha_{s-1} + \alpha_j) \quad \dots (4.3)$$

$$j = 0, \dots, s-1; \quad i = 1, \dots, (s-1)/2$$

are such that in any two positions all combinations of the s elements taken two at a time occur exactly once each. The sets (4.3) written as columns provide $(s(s-1)/2, s, s, 2) : II$, which proves Theorem 2.

For example, if $s = 5$, the residue classes $0, 1, 2, 3, 4 \pmod{5}$ form a field. We write the elements as $0, 1, 2, -1, -2$, and hence the key sets (4.1) are

$$(0, 1, 2, 3, 4)$$

$$(0, 2, 4, 1, 3)$$

the latter being obtained from the former multiplying by 2. Writing them vertically (as shown below in blocks) and generating the other columns by addition of the elements of $GF(5)$ as indicated in (4.3) we obtain the 10 columns

0	1	2	3	4	0	1	2	3	4
1	2	3	4	0	2	3	4	0	1
2	3	4	0	1	4	0	1	2	3
3	4	0	1	2	1	2	3	4	0
4	0	1	2	3	3	4	0	1	2

which is $(10, 5, 5, 2) : II$. This also happens to be $(10, 5, 5, 3) : II$ of strength 3, but a general method of constructing arrays of Type II and strength 3 has to be investigated. Considering only the first three rows we obtain $(10, 3, 5, 2) : II$, which has been used in deriving a cyclic solution to the *combinatorial assignment problem* with 36 officers (Rao, 1961b). Some other applications will be considered in a subsequent communication.

REFERENCES

- BOSE, R. C. (1939) : On the construction of balanced incomplete block designs. *Ann. Eugen.*, **9**, 343-399.
- BOSE, R. C., SHRIKHANDE, S. S. and PARKER, E. T. (1960) : Further results on the construction of mutually Orthogonal Latin Squares and the falsity of Euler's conjecture. *Can. J. Math.*, **12**, 189-203.
- NAIR, K. R. and RAO, C. R. (1948) : Confounding in asymmetrical factorial experiments. *J. Roy. Stat. Soc.*, Suppl. **10**, 109-131.
- PARKER, E. T. (1958) : Construction of some sets of pairwise Orthogonal Latin Squares, Abstract 553-67, *Notices Amer. Math. Soc.*, **5**, 815.
- RAO, C. R. (1946a) : Difference sets and combinatorial arrangements derivable from finite geometries. *Proc. Nat. Inst. Sc.*, **12**, 123-135.
- (1946b) : Hypercubes of strength d leading to confounded designs in factorial experiments. *Bull. Cal. Math. Soc.*, **38**, 67-78.
- (1947) : Factorial experiments derivable from combinatorial arrangements of arrays. *J. Roy. Stat. Soc.*, Suppl., **9**, 128-140.
- (1948) : On a class of arrangements. *Proc. Edin. Math. Soc.*, **8**, 119-125.
- (1956) : A general class of Quasi-factorial designs. *Sankhyā*, **17**, 165-174.
- (1961a) : A study of BIB designs with replications 11 to 15. *Sankhyā*, Series A, **23**, 117-127.
- (1961b) : A combinatorial assignment problem. *Nature*, **191**, 1 July 1961.

Paper received : June, 1961.

ON THE STRUCTURE AND COMBINATORIAL PROPERTIES OF CERTAIN SEMI-REGULAR GROUP DIVISIBLE DESIGNS

By W. S. SARAF

College of Science, Nagpur*

SUMMARY. In this paper, structure of semi-regular group divisible design is investigated and inequalities are found for the number of treatments common to two blocks. Two particular cases of semi-regular group divisible design have been studied in detail and efficient necessary conditions have been deduced for them. One theorem gives a combinatorial property of certain semi-regular group divisible designs. Besides these, there are some minor results.

1. INTRODUCTION

An incomplete design with v treatments, each replicated r times in b blocks, each of size k , forms a group divisible (GD) design if the v treatments can be divided into m groups of n treatments each such that any two treatments belonging to the same group occur together in λ_1 blocks and any two treatments belonging to different groups occur together in λ_2 blocks; $\lambda_1 \neq \lambda_2$. Obviously

$$v = mn, \quad bk = vr \quad \dots (1.1)$$

and

$$r(k-1) = (n-1)\lambda_1 + n(m-1)\lambda_2. \quad \dots (1.2)$$

In this paper we shall study only semi-regular group divisible (SRGD) designs.

For SRGD designs, $(r-\lambda_1) > 0$ and $rk-v\lambda_2 = 0$. Hence (1.2) reduces to

$$r + (n-1)\lambda_1 = n\lambda_2. \quad \dots (1.3)$$

It has been shown by Bose and Connor (1952) that in a SRGD design k is divisible by m and every block contains $c = k/m$ treatments from each group. ... (1.4)

Throughout this paper, in numbering the treatments, we shall adopt the convention that w -th group contains the treatments number

$$(w-1)n+1, (w-1)n+2, \dots, (w-1)n+n = wn; \quad w = 1, 2, \dots, m.$$

Let N denote the Incidence matrix of the design. Then

$$N = \begin{matrix} & \begin{matrix} n_{11} & n_{12} & \dots & n_{1b} \\ n_{21} & n_{22} & \dots & n_{2b} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ n_{v1} & n_{v2} & \dots & n_{vb} \end{matrix} \\ \begin{matrix} (v \times b) \end{matrix} & \end{matrix} \quad \dots (1.5)$$

*Now, District Statistical Officer, Amravati.

Here, the rows correspond to treatments, the columns correspond to blocks and $n_{ij} = 1$ or 0 according as the i -th treatment does or does not occur in the j -th block. It is easy to see that

$$\sum_{j=1}^b n_{ij} = \sum_{j=1}^b n_{ij}^2 = r, \quad \dots (1.6a)$$

$$\sum_{j=1}^b n_{ij} n_{uj} = \lambda_1 \quad \text{or} \quad \lambda_2 \quad (1.6b)$$

according as the i -th and u -th treatments do or do not belong to the same group. Hence, if N' denotes the transpose matrix of N

$$NN' = \begin{bmatrix} A & B & \dots & B \\ B & A & \dots & B \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ B & B & \dots & A \end{bmatrix} \quad \dots (1.7a)$$

where A and B are $n \times n$ matrices defined by

$$A = \begin{bmatrix} r & \lambda_1 & \dots & \lambda_1 \\ \lambda_1 & r & \dots & \lambda_1 \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \lambda_1 & \lambda_1 & \dots & r \end{bmatrix}; \quad B = \begin{bmatrix} \lambda_2 & \lambda_2 & \dots & \lambda_2 \\ \lambda_2 & \lambda_2 & \dots & \lambda_2 \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \lambda_2 & \lambda_2 & \dots & \lambda_2 \end{bmatrix} \quad \dots (1.7b)$$

2. GENERAL BOUNDS ON THE NUMBER OF TREATMENTS COMMON TO TWO BLOCKS

Choose any $t \leq b$ blocks of the design. Let the sub-matrix of N that corresponds to these t blocks be denoted by N_0 and let s_{ju} be the number of treatments common to the j -th and u -th blocks ($j, u = 1, 2, \dots, t$). Then the $t \times t$ matrix

$$N'_0 N_0 = S_t = (s_{ju}) \quad \dots (2.1)$$

is the structural matrix of the t chosen blocks. Adjoin m columns to the incidence matrix N so that the i -th adjoined column consists of 1's in the rows that correspond to the treatments of the i -th group and 0's elsewhere; $i = 1, 2, \dots, m$. If this augmented matrix be denoted by N_1 then

$$N_1 N'_1 = \begin{bmatrix} D & B & \dots & B \\ B & D & \dots & B \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ B & B & \dots & D \end{bmatrix} \quad \dots (2.2a)$$

STRUCTURE AND COMBINATORIAL PROPERTIES OF SRGD DESIGNS

where D is an $n \times n$ matrix defined by

$$D = \begin{bmatrix} r+1 & \lambda_1+1 & \dots & \lambda_1+1 \\ \lambda_1+1 & r+1 & \dots & \lambda_1+1 \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \lambda_1+1 & \lambda_1+1 & \dots & r+1 \end{bmatrix}. \quad \dots (2.2b)$$

Using (1.1) and (1.3) we get

$$|N_1 N_1'| = (rk+n)n^{m-1}(r-\lambda_1)^{m(n-1)} \quad \dots (2.2c)$$

which is always non-vanishing for SRGD design.

Let the columns of N_1 be permuted so that the first t columns correspond to the chosen blocks; then the sub-matrix of N_1 that corresponds to these t blocks will be N_0 . Now, adjoin t row vectors to N_1 such that the i -th row consists of zero elements everywhere except the i -th which is unity. If this resultant matrix of order $(v+t) \times (b+m)$ be denoted by N_2 then

$$N_2 = \begin{bmatrix} N_1 \\ \hline I_t \quad 0 \end{bmatrix} \quad \dots (2.3)$$

where I_t is the identity matrix of order t and 0 is a $t \times (b+m-t)$ matrix with all null elements. Further

$$N_2 N_2' = \begin{bmatrix} N_1 N_1' & N_0 \\ \hline N_0' & I_t \end{bmatrix} \quad \dots (2.4a)$$

and

$$|N_2 N_2'| = (n+v\lambda_2)^{-t+1} n^{m-1}(r-\lambda_1)^{v-t-m} |C_t| \quad \dots (2.4b)$$

where C_t is of order $t \times t$ and its elements are

$$C_{ju} = (n+v\lambda_2)\delta_{ju} + kc(\lambda_1+1-\lambda_2) + k^2\lambda_2 \quad \dots (2.5)$$

where $\delta_{ju} = (r-\lambda_1-k)$ or $-s_{ju}$ according as $j = u$ or $j \neq u$. The matrix C_t given by (2.4b) is defined as the characteristic matrix of the t chosen blocks. It is a symmetric matrix. If S_t is the structural matrix of the t blocks then there exists a one-to-one correspondence between the elements of C_t and S_t . The relationship between the matrices C_t and S_t can be expressed as

$$C_t = (n+v\lambda_2)(r-\lambda_1)I_t - (n+v\lambda_2)S_t + [kc(\lambda_1+1-\lambda_2) + k^2\lambda_2]E_t \quad \dots (2.6)$$

where E_t is a $t \times t$ matrix with all elements unity.

We can now state the following.

Theorem 1: If C_t is the characteristic matrix of any set of t blocks chosen from a SRGD design with parameters $v, b, r, k, \lambda_1, \lambda_2, m, n$, then (i) $|C_t| \geq 0$ if $t < b+m-v$; (ii) $|C_t| = 0$ if $t > b+m-v$ and (iii) $(n+v\lambda_2)^{-t+1} n^{m-1} (r-\lambda_1)^{v-t-m} |C_t|$ is a perfect square if $t = b+m-v$.

As a simple consequence of Theorem 1, we get the following inequalities for s_{ju} . Let $t = 2$. Since the factor outside $|C_t|$ in (2.4b) is positive,

$$|C_2| = [(n+v\lambda_2)(r-\lambda_1-k) + kc(\lambda_1+1-\lambda_2) + k^2\lambda_2]^2 - [-(n+v\lambda_2)s_{12} + kc(\lambda_1+1-\lambda_2) + k^2\lambda_2]^2 \geq 0 \dots (2.7)$$

Simplifying this, we get the following:

Corollary 1: For a SRGD design, the number of treatments common to any two blocks satisfies the inequalities

$$r-\lambda_1-k + \frac{2k}{n+v\lambda_2} [c(\lambda_1+1-\lambda_2) + k\lambda_2] \geq s_{ju} \geq -(r-\lambda_1-k).$$

For symmetric SRGD design, i.e., the design with $r = k$, s_{ju} satisfies

$$-\lambda_1 + \frac{2k}{n+v\lambda_2} [c(\lambda_1+1-\lambda_2) + k\lambda_2] \geq s_{ju} \geq \lambda_1. \dots (2.8)$$

Connor (1952) has proved that for a symmetric SRGD design s_{ju} satisfies the inequalities

$$-\lambda_1 + \frac{2\lambda_2 r^2}{r+v\lambda_2-\lambda_1} \geq s_{ju} \geq \lambda_1. \dots (2.9)$$

It is easy to see that (2.8) gives stronger bounds for s_{ju} than (2.9). When $(\lambda_2-\lambda_1) \neq 1$, (2.8) and (2.9) are identical.

3. SRGD OF LINKED BLOCK TYPE

Roy and Laha (1956) have proved that for a LB type SRGD design $b = v-m+1$. Consider the characteristic matrix of $t = 2$ blocks of this design. Obviously $t > b+m-v$ and hence by (ii) of Theorem 1,

$$|C_2| = [(n+v\lambda_2)(r-\lambda_1-k) + kc(\lambda_1+1-\lambda_2) + k^2\lambda_2]^2 - [-(n+v\lambda_2)s_{12} + kc(\lambda_1+1-\lambda_2) + k^2\lambda_2]^2 = 0. \dots (3.1)$$

$$\text{or } (n+v\lambda_2)(r-\lambda_1-k) + kc(\lambda_1+1-\lambda_2) + k^2\lambda_2 = \pm [-(n+v\lambda_2)s_{12} + kc(\lambda_1+1-\lambda_2) + k^2\lambda_2]. \dots (3.2)$$

STRUCTURE AND COMBINATORIAL PROPERTIES OF SRGD DESIGNS

The negative sign on the right hand side of (3.2) is discarded since it gives fractional value of s_{12} which is essentially an integer. Prefixing positive sign to the right hand side of (3.2), we get

$$s_{12} = k + \lambda_1 - r. \quad \dots (3.3)$$

This important result which determines the structure of this design completely is stated in the following theorem.

Theorem 2 : *For a SRGD design with $b = v - m + 1$, any two blocks have $(k + \lambda_1 - r)$ treatments in common.*

From this theorem, it is obvious that when a SRGD design is dualised, the dual design is a BIB design. Hence the following corollary.

Corollary 2 : *If an LB type SRGD design with parameters $v, b, r, k, m, n, \lambda_1, \lambda_2$ is dualised, we get a BIB design with the following parameters:*

$$v^* = b, b^* = v, r^* = k, k^* = r, \lambda^* = k + \lambda_1 - r.$$

The following two corollaries are also obvious.

Corollary 3 : *In an LB type SRGD design, if one block is omitted and in the remaining blocks only those treatments are retained that occur in the omitted block, then a GD design with following parameters is obtained.*

$$b^* = b - 1, v^* = k, r^* = r - 1, k^* = k + \lambda_1 - r, m^* = m, n^* = c, \lambda_1^* = \lambda_1 - 1, \lambda_2^* = \lambda_2 - 1.$$

Corollary 4 : *If one block and all treatments occurring in it are omitted from an LB type SRGD design then a GD design with the following parameters is obtained.*

$$b^* = b - 1, v^* = v - k, r^* = r, k^* = r - \lambda_1, m^* = m, n^* = n - c, \lambda_1^* = \lambda_1, \lambda_2^* = \lambda_2.$$

From Theorem 2, we can obtain a necessary condition for the existence of SRGD design with $b = v - m + 1$.

To the incidence matrix N , adjoin $(m-1)$ new columns such that the i -th adjoined column has 1's in the positions $(i-1)n+1, (i-1)n+2, \dots (i-1)n+n$ and 0's elsewhere. If this resultant square matrix of order v is denoted by A , then using Theorem 2 we have :

$$A'A = \begin{bmatrix} k & (k+\lambda_1-r) & \dots & (k+\lambda_1-r) & c & c & \dots & c \\ (k+\lambda_1-r) & k & & \dots & (k+\lambda_1-r) & c & c & \dots & c \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ (k+\lambda_1-r) & (k+\lambda_1-r) & \dots & k & c & c & \dots & c \\ c & c & \dots & c & n & 0 & \dots & 0 \\ c & c & \dots & c & 0 & n & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ c & c & \dots & c & 0 & 0 & \dots & n \end{bmatrix} \quad \dots (3.4a)$$

and

$$|A'A| = |A|^2 = n^m \lambda_2 (r - \lambda_1)^{b-1} = n^v \lambda_2 (\lambda_2 - \lambda_1)^{b-1} \quad \dots (3.4b)$$

which must be perfect integral square. This leads to the following theorem.

Theorem 3 : (i) *With even number of treatments in an LB type SRGD design, λ_2 or $\lambda_2(\lambda_2 - \lambda_1)$ must be perfect integral square according as b is odd or even.*

(ii) *With odd number of treatments in an LB type SRGD design $n\lambda_2$ is a perfect square.*

As an illustration, consider the design with $v = 93$, $b = 63$, $k = 31$, $r = 21$, $m = 31$, $n = 3$, $\lambda_1 = 0$, $\lambda_2 = 7$ as the parameters. For this design $n\lambda_2 = 21$. Hence according to (ii) of Theorem 3 the design is impossible. For an LB type SRGD design with $\lambda_1 = 0$, $\lambda_2 = 1$ any two blocks have exactly one treatment in common. Therefore, by Theorem 2, substituting $\lambda_1 = 0$

$$k - r = 1 \quad \text{or} \quad r = k - 1. \quad \dots (3.5)$$

Hence, we have proved the following lemma.

Lemma 1 : *For an LB type SRGD design with $\lambda_1 = 0$, $\lambda_2 = 1$, $r = k - 1$. Using (1.1), (1.3) and the Lemma 1, it is easy to see that the parameters of an LB type SRGD design with $\lambda_1 = 0$ and $\lambda_2 = 1$ can be expressed as:*

$$v = m(m-1); \quad b = (m-1)^2, \quad r = (m-1), \quad k = m, \quad n = m-1, \quad m, \quad \lambda_1 = 0, \quad \lambda_2 = 1.$$

Majumdar (1953) has shown that a symmetric BIB design is the only design in which any two varieties occur together in λ blocks and any two blocks have λ' varieties ($\lambda = \lambda'$) in common. A somewhat similar property holds good for an LB type SRGD design. It is stated as follows:

If in an incomplete design, $v = mn$ treatments divided into m groups of n treatments each, are applied in b blocks such that any two treatments belonging to the same group occur together in λ_1 blocks, any two treatments belonging to different groups occur together in λ_2 blocks ($\lambda_1 < \lambda_2$) and any two blocks have $\lambda' (\neq 0)$ treatments in common, then the design is an LB type SRGD design.

The proof is implicitly contained in Roy and Laha (1956).

4. AFFINE RESOLVABILITY OF SRGD DESIGN

In a resolvable design, the blocks are separable into groups, each group (or replication) containing all varieties, each variety once and only once. Each replication must necessarily contain the same number of blocks, say t , so that,

$$b = tr, \quad v = tk. \quad \dots (4.1)$$

We shall now prove the following theorem.

Theorem 4 : *For a resolvable SRGD design with $b = v - m + r$, any two blocks not belonging to the same replication have the same number of treatments in common.*

STRUCTURE AND COMBINATORIAL PROPERTIES OF SRGD DESIGNS

Let B_{ij} denote the i -th block in the j -th replication. $i = 1, 2, \dots, t$; $j = 0, 1, 2, \dots, r-1$, and l_{ij} the number of common treatments in the blocks B_{10} and B_{ij} ; $i = 1, 2, \dots, t$; $j = 1, 2, \dots, r-1$. Then

$$\sum_{i=1}^t \sum_{j=1}^{r-1} l_{ij} = k(r-1). \quad \dots (4.2)$$

Considering the combinations of varieties taken two at a time from the block B_{10} :

$$\sum_{j=1}^{r-1} \sum_{i=1}^t \frac{l_{ij}(l_{ij}-1)}{2} = \frac{mc(c-1)(\lambda_1-1)}{2} + \frac{m(m-1)c^2(\lambda_2-1)}{2} \quad \dots (4.3)$$

$$\sum_{j=1}^{r-1} \sum_{i=1}^t l_{ij}^2 = k(c-1)(\lambda_1-1) + (\lambda_2-1)m(m-1)c^2 + k(r-1). \quad \dots (4.4)$$

Let $\bar{x}$ denote the mean and s^2 be the mean sum of squares for the $t(r-1)$ quantities l_{ij} . $i = 1, 2, \dots, t$; $j = 1, 2, \dots, (r-1)$. Then

$$\bar{x} = \frac{\sum_{j=1}^{r-1} \sum_{i=1}^t l_{ij}}{t(r-1)} = \frac{k}{t} \quad \dots (4.5)$$

and

$$s^2 = \frac{1}{t(r-1)} \sum_{j=1}^{r-1} \sum_{i=1}^t l_{ij}^2 - \bar{x}^2. \quad \dots (4.6)$$

From (4.4)
$$\sum_{j=1}^{r-1} \sum_{i=1}^t l_{ij}^2 = k(c-1)(\lambda_1-1) + kc(m-1)(\lambda_2-1) + k(r-1).$$

$$= k(r-\lambda_1) + kc(\lambda_1-\lambda_2) + k^2(\lambda_2-1) \quad \dots (4.7)$$

$$\therefore t(r-1)s^2 = k(r-\lambda_1) + kc(\lambda_1-\lambda_2) + k^2(\lambda_2-1) - \frac{k^2}{t^2} \cdot t(r-1). \quad \dots (4.8)$$

This simplifies to
$$t(r-1)s^2 = \frac{k}{v}(v-k)(r-\lambda_1-k). \quad \dots (4.9)$$

But $s^2 \geq 0$ and since $v > k$ we have

$$r \geq k + \lambda_1. \quad \dots (4.10)$$

When the equality holds good, i.e., when $r = k + \lambda_1$, then $s^2 = 0$. Hence, each $l_{ij} = \bar{x}$ or k/t . Such designs may be called "affine resolvable SRGD designs." Now for affine resolvable SRGD design,

$$r = k + \lambda_1 = \frac{vr}{b} + \lambda_1 \quad \dots (4.11)$$

or

$$b(r-\lambda_1) = vr. \quad \dots (4.12)$$

Since $r - \lambda_1 > 0$, this simplifies to

$$b = v - m + r. \quad \dots (4.13)$$

Hence if the design is resolvable then if $b = v - m + r$, it is affine resolvable and conversely. This proves the theorem. Also, it is clear that for affine resolvable design, k/t must be an integer.

With the help of the above theorem, a necessary condition for the existence of the affine resolvable SRGD design can be deduced.

In forming the incidence matrix N for a resolvable SRGD design the blocks shall be arranged in such a way that the first t columns correspond to the blocks of the first replication; next t columns to those of the second replication and so on.

To the incidence matrix N adjoin m columns such that the j -th adjoined column contains 1's in the positions $(j-1)n+1, (j-1)n+2, (j-1)n+3, \dots, (j-1)n+n = jn$, and 0's elsewhere. Then, this new matrix is extended by joining r rows such that the i -th adjoined row will have 1's in the positions $(i-1)t+1, (i-1)t+2, (i-1)t+3, \dots, (i-1)t+t = it$, and 0's elsewhere.

If this augmented square matrix of order $(v+r)$ be denoted by N^* then

$$N^*N^{*'} = \begin{bmatrix} r+1 & \lambda_1+1 & \dots & \lambda_1+1 & \lambda_2 & \dots & \dots & \lambda_2 & 1 & 1 & \dots & 1 \\ \lambda_1+1 & r+1 & \dots & \lambda_1+1 & \lambda_2 & \dots & \dots & \lambda_2 & 1 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \lambda_1+1 & \dots & \dots & r+1 & \lambda_2 & \dots & \dots & \lambda_2 & 1 & 1 & \dots & 1 \\ \lambda_2 & \dots & \dots & \lambda_2 & r+1 & \lambda_1+1 & \lambda_2 & \lambda_2 & 1 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \lambda_2 & \dots & \dots & \lambda_2 & \lambda_1+1 & r+1 & \lambda_2 & \lambda_2 & 1 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \lambda_2 & \dots & \dots & \dots & \dots & \dots & \lambda_1+1 & r+1 & 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & \dots & \dots & \dots & \dots & \dots & 1 & t & 0 & 0 \\ 1 & 1 & \dots & \dots & \dots & \dots & \dots & \dots & 1 & 0 & t & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & \dots & \dots & \dots & \dots & \dots & 1 & 0 & 0 & t \end{bmatrix} \quad \dots (4.14)$$

Using (1.1), (1.3) and $rk - v\lambda_2 = 0$

$$|N^*N^{*'}| = |N^*|^2 = t^r n^m k^{m(n-1)} = v^r \cdot n^m \cdot k^{b-2r} \quad \dots (4.15)$$

since

$$b = v - m + r \text{ and } tk = v.$$

(4.15) must be a perfect square.

We can now state the following theorem.

Theorem 5: (i) *With even number of blocks in an affine resolvable SRGD design (a) m must be a perfect square if m and r are odd. (b) n must be a perfect square if m is odd and r is even.*

STRUCTURE AND COMBINATORIAL PROPERTIES OF SRGD DESIGNS

(ii) With odd number of blocks in an affine resolvable SRGD design (a) c must be a perfect square if m and r are odd. (b) nk must be a perfect square if m is odd and r is even. (c) t must be a perfect square if m is even and r is odd.

5. COMBINATORIAL PROPERTY FOR CERTAIN SRGD DESIGNS

We shall conclude this paper after proving a combinatorial property for certain SRGD designs. It is stated in the following :

Theorem 6 : *The necessary and sufficient condition for the existence, in a SRGD design, of two blocks such that the number of treatments common to any other block and these two are equal, is that the two blocks contain $(k + \lambda_1 - r)$ common treatments.*

The proof is almost on the same lines as given by Majumdar (1953) for a similar property of a BIB design.

Let the two blocks which may be called the first and second blocks have s treatments in common. Of these s treatments, let there be l_j treatments from the j -th group. $j = 1, 2, \dots, m$.

Let the i -th block have x_{ij} treatments from j -th group that occur in the first block but not in the second; and y_{ij} treatments from j -th group that occur in the second but not in the first block.

Then, considering the combinations of treatments, taken two at a time,

$$\begin{aligned} \sum_{i=3}^b \sum_{j=1}^m x_{ij} (x_{ij} - 1) + \sum_{i=3}^b \sum_{\substack{j, j'=1 \\ j \neq j'}}^m x_{ij} x_{ij'} &= \sum_{i=3}^b \sum_{j=1}^m y_{ij} (y_{ij} - 1) + \sum_{i=3}^b \sum_{\substack{j, j'=1 \\ j \neq j'}}^m y_{ij} y_{ij'} \\ &= \sum_{j=1}^m (c - l_j)(c - l_j - 1)(\lambda_1 - 1) + \sum_{\substack{j, j'=1 \\ j \neq j'}}^m (c - l_j)(c - l_{j'}) (\lambda_2 - 1) \end{aligned} \quad \dots (5.1)$$

$$\sum_{i=3}^b \sum_{j=1}^m x_{ij} y_{ij} = \sum_{j=1}^m (c - l_j)^2 \lambda_1. \quad \dots (5.2)$$

$$\sum_{i=3}^b \sum_{\substack{j, j'=1 \\ j \neq j'}}^m x_{ij} y_{ij'} = \sum_{\substack{j, j'=1 \\ j \neq j'}}^m (c - l_j)(c - l_{j'}) \lambda_2. \quad \dots (5.3)$$

$$\sum_{i=3}^b \sum_{j=1}^m x_{ij} = \sum_{i=3}^b \sum_{j=1}^m y_{ij} = \sum_{j=1}^m (c - l_j)(r - 1) \quad \dots (5.4)$$

using these we have

$$\sum_{i=3}^b \left[\sum_{j=1}^m x_{ij} - \sum_{j=1}^m y_{ij} \right]^2 = 2(k - s)(r - \lambda_1 - k - s).$$

Hence, if

$$s = k + \lambda_1 - r, \quad \sum_{j=1}^m x_{ij} = \sum_{j=1}^m y_{ij}$$

and conversely. This proves the theorem.

I wish to acknowledge my sincere thanks to Professor S. S. Shrikhande, who verified the proofs and suggested important corrections.

REFERENCES

- BOSE, R. C. and CONNOR, W. S. (1952): Combinatorial properties of group divisible incomplete block designs. *Ann. Math. Stat.*, **23**, 367-386.
- CONNOR, W. S. (1952): On the structure of balanced incomplete block designs. *Ann. Math. Stat.*, **23**, 57-71.
- (1952): Some relations among the blocks of symmetrical group divisible design. *Ann. Math. Stat.*, **23**, 602-609.
- MAJUMDAR, K. N. (1953): On some theorems in combinatorics relating to incomplete block designs. *Ann. Math. Stat.*, **24**, 377-389.
- ROY, J. and LAHA, R. G. (1956): Classification and analysis of linked block designs. *Sankhyā*, **17**, 115-132.
- SHRIKHANDE, S. S. (1953): The non-existence of certain affine resolvable balanced incomplete block designs. *Canadian Journal of Mathematics*, **5**, 413-420.

Paper received : October, 1959.

Revised : October, 1960.

MISCELLANEOUS

A FACTORIAL APPROACH TO CONSTRUCTION OF TRUE COST OF LIVING INDEX AND ITS APPLICATION IN STUDIES OF CHANGES IN NATIONAL INCOME

By K. S. BANERJEE

State Statistical Bureau, West Bengal

SUMMARY. A new formula for the *true index* of cost of living has been evolved through a new formulation. Incidentally, it has been shown how the price and the quantity components of value change in National Income can be evaluated.

1. INTRODUCTION

A new formula has been evolved for the *true index* of cost of living through a factorial approach and this has been shown to possess all the properties of Wald's true index (1939) as well as those of Frisch's true index (1936) obtained by the double-expenditure method especially in so far as it lies between Staehle's limits (1935).

2. NEW FORMULATION

An indifference-defined price index denotes the ratio of expenditures between two situations ensuring the same standard of living. Similarly, a quantity index also stands for the ratio of the two standards of living. As two points on an indifference surface are *equivalent*, the quantity index between such a pair will be unity. Conversely, if the quantity index be equal to unity, it implies that the standards of living in the two situations are the same. Moreover, the quantity index between two equivalent points q_1 and q_0 on $I(q)$ will also be unity where $I(q)$ stands for the choice indicator describing the scale of preference of an individual. If, therefore, a quantity q_1 could be found (by any formulation), in situation 1 such that the quantity index κ_{01} , between q_0 and q_1 , is unity, then q_1 can be regarded as equivalent to q_0 . The price index, π_{01} , found by such a joint formulation, will represent the true index of cost of living of the period 1 compared to the period 0.

As an illustration of our formulation, we may consider Fisher's ideal formulae for price and quantity :

$$F_p = \sqrt{\frac{V_{10}}{V_{00}} \frac{V_{11}}{V_{01}}} \quad \dots (2.1)$$

$$\text{and} \quad F_q = \sqrt{\frac{V_{01}}{V_{00}} \frac{V_{11}}{V_{10}}} \quad \dots (2.2)$$

$$\text{where} \quad V_{jk} = \sum_{i=1}^n p_j^i q_k^i \quad (j, k = 0, 1) \quad \dots (2.3)$$

is the value of $q_k^i = (q_k^1, q_k^2, \dots, q_k^n)$ calculated at prices $p_j^i = (p_j^1, p_j^2, \dots, p_j^n)$; F_p and F_q are Fisher's ideal indexes for price and quantity respectively. If now $F_q = 1$, the price index reduces to V_{11}/V_{00} .

The condition $F_q = 1$ implies

$$V_{11} V_{01} = V_{10} V_{00}, \quad \dots (2.4)$$

a condition which was independently worked out by Frisch in his double-expenditure approach. (Frisch, 1936).

3. FACTORIAL APPROACH

Let p_0^i and q_0^i be the price and quantity of the i -th commodity in the situation 0, and let p_1^i and q_1^i be the corresponding figures in the situation 1. From these two sets of entries we obtain four values for the i -th item: V_{00}^i , V_{01}^i , V_{10}^i , and V_{11}^i where, of course, $V_{jk}^i = p_j^i q_k^i$ ($j, k = 0, 1$). In analogy with a factorial experiment involving two factors at two levels each, we have the price and quantity components as follows:

$$P^i = \frac{1}{2}[V_{11}^i - V_{01}^i + V_{10}^i - V_{00}^i] = \frac{1}{2}V_{00}^i(\pi^i - 1)(\kappa^i + 1) \quad \dots (3.1)$$

$$Q^i = \frac{1}{2}[V_{11}^i - V_{10}^i + V_{01}^i - V_{00}^i] = \frac{1}{2}V_{00}^i(\pi^i + 1)(\kappa^i - 1), \quad \dots (3.2)$$

where $\pi^i = \frac{p_1^i}{p_0^i}$ and $\kappa^i = \frac{q_1^i}{q_0^i}$ are the price and quantity indexes respectively of the i -th item. It is easy to see that the value change, $V_{11}^i - V_{00}^i$, can simply be written as the sum of the price and quantity components. That is,

$$V_{11}^i - V_{00}^i = P^i + Q^i, \quad \dots (3.3)$$

which when summed over all the n commodities yields the following:

$$V_{11} - V_{00} = P + Q, \quad \dots (3.4)$$

where $P = \sum_{i=1}^n P^i$, $Q = \sum_{i=1}^n Q^i$ and $V_{jj} = \sum_{i=1}^n V_{jj}^i$, ($j = 0, 1$). ... (3.5)

If now we define π and κ as the *general* price and quantity indexes respectively we can, by analogy of identities (3.1) and (3.2), write

$$P = \frac{1}{2}V_{00}(\pi - 1)(\kappa + 1), \quad \dots (3.6)$$

and $Q = \frac{1}{2}V_{00}(\pi + 1)(\kappa - 1). \quad \dots (3.7)$

Following Stuvell (1957) we may equate these to ΣP^i and ΣQ^i , where P^i and Q^i are defined by (3.1) and (3.2). That is,

$$\frac{1}{2} V_{00}(\pi - 1)(\kappa + 1) = \frac{1}{2}(V_{11} - V_{01} + V_{10} - V_{00}) \quad \dots (3.8)$$

$$\frac{1}{2} V_{00}(\pi + 1)(\kappa - 1) = \frac{1}{2}(V_{11} - V_{10} + V_{01} - V_{00}) \quad \dots (3.9)$$

or, alternatively, after removing the factor $\frac{1}{2}V_{00}$ from both sides,

$$(\pi - 1)(\kappa + 1) = \frac{V_{11}}{V_{00}} - L_q + L_p - 1 \quad \dots (3.10)$$

$$(\pi + 1)(\kappa - 1) = \frac{V_{11}}{V_{00}} - L_p + L_q - 1, \quad \dots (3.11)$$

ON THE CONSTRUCTION OF TRUE COST OF LIVING INDEX

where $\frac{V_{11}}{V_{00}}$ is the value index, L_p and L_q are the Laspeyres' price and quantity indexes, i.e. $L_p = V_{10}/V_{00}$, $L_q = V_{01}/V_{00}$. From (3.8) and (3.9) we have :

$$\pi - k = (L_p - L_q) \quad \dots (3.12)$$

$$\pi k = \frac{V_{11}}{V_{00}} \quad \dots (3.13)$$

from which the price and quantity indexes are obtained as

$$\pi = \frac{1}{2}(L_p - L_q) + \sqrt{\left[\frac{1}{2}(L_p - L_q)\right]^2 + \frac{V_{11}}{V_{00}}} \quad \dots (3.14)$$

$$k = \frac{1}{2}(L_q - L_p) + \sqrt{\left[\frac{1}{2}(L_p - L_q)\right]^2 + \frac{V_{11}}{V_{00}}} \quad \dots (3.15)$$

On account of the symmetry in the relationship of π and k in (3.14) and (3.15), the indexes π and k satisfy certain fundamental tests, such as the time reversal and the factor reversal tests and are, therefore, preferred to other known formulae (Stuvel).

The condition of equivalence in our formulation can be worked out by equating the quantity index to unity. If this were done, i.e., if $\kappa = 1$, it would follow from (3.11) that the quantity component in the value change is zero, i.e.,

$$V_{11} - V_{10} + V_{01} - V_{00} = 0 \quad \dots (3.16)$$

and also from (3.13),
$$\pi = \frac{V_{11}}{V_{00}}. \quad \dots (3.17)$$

Relation (3.16) provides the necessary condition for determining the point q_1 such that it is equivalent to q_0 . This condition has a striking similarity with that of Frisch. In the latter case, however, the condition may be written in terms of logarithms, i.e.,

$$\log V_{11} - \log V_{10} + \log V_{01} - \log V_{00} = 0 \quad \dots (3.18)$$

which is, in fact, similar to (3.16).

The condition of equivalence could also be worked out from formulae other than Stuvel's. In fact such conditions have already been derived in the case of Fisher's ideal index which is rather a special case of the generalised indexes of Stuvel (Banerjee). In particular cases of the generalised index, however, the conditions of equivalence take complicated expressions.

4. THE NEW INDEX COMPARED WITH THOSE OF WALD AND FRISCH

We next compare our index obtained in (3.17) with Frisch's true index as also with Wald's true index. Suppose the Engel curves C_0 and C_1 are linear and are given by the following equations.

The equations of

$$\begin{aligned} C_0 : q_0^i &= \alpha_0^i V + \beta_0^i \\ C_1 : q_1^i &= \alpha_1^i V + \beta_1^i \\ (i &= 1, 2, \dots, n). \end{aligned} \quad \dots (4.1)$$

Let $\sum_{i=1}^n \alpha_j^i p_k^i = a_{jk}$ and $\sum_{i=1}^n \beta_j^i p_k^i = b_{jk}$, ($j, k = 0, 1$). ... (4.2)

Then

$$a_{jj} = 1, \quad b_{jj} = 0 \quad (j = 0, 1).$$

Also

$$\sum_{i=1}^n p_0^i q_0^i = V_{00}, \quad V_{01} = a_{10} V_{11} + b_{10}, \quad V_{10} = a_{01} V_{00} + b_{01}.$$

In terms of the above coefficients, Frisch's index is given by

$$F = \frac{-b_{10} + \sqrt{b_{10}^2 + 4a_{10}a_{01}V_{00}^2 + 4a_{10}b_{01}V_{00}}}{2a_{10}V_{00}} \quad \dots (4.3)$$

and, Wald's true index is given by

$$W = \sqrt{\frac{a_{01}}{a_{10}}} + \frac{1}{V_{00}} \frac{b_{01} - \frac{b_{10}}{a_{10}} \sqrt{a_{01}a_{10}}}{1 + \sqrt{\frac{a_{01}}{a_{10}}}} \quad \dots (4.4)$$

while our index, obtained through the condition (3.16) of equivalence, is obtained as

$$B = \frac{1+a_{01}}{1+a_{10}} + \frac{1}{V_{00}} \frac{(b_{01}-b_{10})}{1+a_{10}}. \quad \dots (4.5)$$

Following the same line of argument, as adopted by Wald, it can be shown that our index B also lies between the limits of Staehle (see Wald).

That is,

$$\frac{V_{11}}{V_{01}} \leq B \leq \frac{V_{10}}{V_{00}}, \quad \dots (4.6)$$

where $V_{01} = V_{00}$. Or, in terms of the coefficients of the Engel curves, the new index lies between,

$$\frac{1}{a_{10}} - \frac{1}{V_{00}} \frac{b_{10}}{a_{10}} \quad \text{and} \quad a_{01} + \frac{b_{01}}{V_{00}}. \quad \dots (4.7)$$

When, however, $a_{10} = a_{01}$, we see $W = B$. If, in addition, $b_{10} = b_{01}$, we see $F = W = B = 1$.

Although formulae (4.3)–(4.5) are exact in terms of the coefficients of the Engel curves, yet in practice, the indexes have to be computed from *estimates* of the coefficients of the Engel curves which are subject to errors of estimation. The error of these indexes can, of course, be determined taking p_0 and p_1 as given, since the standard errors of the regression coefficients are known.

ON THE CONSTRUCTION OF TRUE COST OF LIVING INDEX

Even if the point q_1 is such that Q , i.e., the quantity component in the value change, is *not* exactly equal to zero, the price formula given in (3.14) may still be considered as a good approximation to the true index, provided the quantity component in the value change did not differ from zero significantly. Tests of significance similar to those applied to factorial experiments may perhaps be applied in our case also, though with caution, in view of non-experimental nature of economic data.

In the next section we consider a practical situation for which our methods could be applied with some advantages in explaining value changes in terms of price and quantity variations.

5. PRICE AND QUANTITY COMPONENTS IN VALUE CHANGE OF NATIONAL INCOME

Estimates of national income are available for Indian Union in a series of papers entitled, "Estimates of National Income", published by the Central Statistical Organisation. These papers show, besides other particulars, a statement on the movement of net national output at factor cost from year to year. Comparison of the national output is made possible, as usual, through the indexes of national output provided at current prices (indexes of value) and constant prices (indexes of output or quantity). These indexes furnish, of course, a basis of comparison, but a complete picture of the movement of national income will not be available unless it is known how much of the change in value is due to change in price and how much to change in output (quantity).

We shall now apply formulae (3.4)–(3.7) to Indian national income data to explain changes in the national income in terms of changes in prices and also changes in the volume of national output.

In "Estimates of National Income", the values of L_q (quantity index) and V_{11}/V_{00} (value index) are provided, but not the values of L_p (price index). For L_p (Laspeyres' price index), we may take the wholesale price indexes* for India perhaps as close substitutes. These price indexes (base 1939 = 100) are not, of course, identical with the price indexes which would have been obtained from the statistics utilised in the estimation of national income. However, the order of disagreement between the two, is probably not so large as to vitiate the study to an unacceptable extent, or to make the computations absolutely hypothetical.

The computed price components and quantity components in value change are given in the following Table for the period 1949-50 to 1957-58. In the Table, the values of L_q (Col. 3) and V_{11}/V_{00} (Col. 4) have been taken from the papers on national income, while, the values of L_p (Col. 2), as stated before, are the wholesale price indexes obtained as simple averages of the monthly indexes for 12 months of the corresponding fiscal years; π and k (Cols. 6 and 7 respectively) are the adjusted price and quantity indexes. Column 5 shows the percentage change of value of net national output as given by

$$100 \left(\frac{V_{11}}{V_{00}} - 1 \right). \quad \dots (5.1)$$

* Vide reports issued by the Office of the Economic Adviser, Government of India.

SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS : SERIES A

TABLE 1. PRICE AND QUANTITY COMPONENTS

years	L_p	L_q	$100V_{11}/V_{00}$	$100(V_{11}/V_{00}-1)$	π	κ	P	Q
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1948-49	100.0	100.0	100.0	0.0	100.0	100.0	0.0	0.0
-50	103.0	102.0	104.2	4.2	102.6	101.6	2.6	1.6
-51	109.6	102.3	110.2	10.2	108.7	101.4	8.7	1.5
-52	116.3	105.2	115.3	15.3	113.1	102.0	13.2	2.1
-53	101.8	109.4	113.5	13.5	102.8	110.4	3.0	10.5
-54	106.3	116.0	121.2	21.2	105.4	115.1	5.8	15.4
-55	101.0	118.8	111.1	11.1	96.4	114.2	-3.4	14.5
-56	96.4	121.2	115.4	15.4	95.7	120.5	-4.7	20.1
-57	107.1	127.2	130.8	30.8	104.8	124.9	5.3	25.5
-58	110.3	125.2	131.3	31.3	107.4	122.3	8.2	23.1

The percentage difference as shown in Col. 5 is obviously the sum of the components in Cols. 8 and 9.

Table 1 shows that during the years, 1954-55 and 1955-56, the price component, that is, the contribution made by price to value change, has been negative, as the corresponding price index (adjusted) is less than 100. As a result, a part of the contribution made by quantity (output) to the increase in value of national income during those years has been masked by the depressing price effect.

6. GENERALISATION OF TWO-FACTOR INDEX NUMBERS

Index numbers for price (p) and quantity (q) involving two factors may be regarded as two-dimensional, as only two factors, price and quantity, are involved, e.g., Laspeyres', Paasche's or Fisher's index. Multidimensional index numbers involving more than two factors could also be constructed by a direct generalisation of (3.14) and (3.15). For example, if p, q, r be the factors observed each at two levels 0 and 1, the three dimensional index for p may be written as

$$\pi = \frac{1}{2}(L_{p/qr} - L_{qr/p}) + \sqrt{[\frac{1}{2}(L_{p/qr} - L_{qr/p})]^2 + R_{pqr}} \quad \dots (6.1)$$

ON THE CONSTRUCTION OF TRUE COST OF LIVING INDEX

where

$$\begin{aligned} L_{p|qr} &= V_{100}/V_{000}, & L_{qr|p} &= V_{011}/V_{000} \\ R_{pqr} &= V_{111}/V_{000}, \end{aligned} \quad \dots \quad (6.2)$$

the V 's being defined by
$$V_{jkl} = \sum_{i=1}^n p_i^j q_i^k r_i^l, \quad (j, k, l = 0, 1). \quad \dots \quad (6.3)$$

Similarly, three dimensional indexes for q and r can be defined. A mention may perhaps be made of the work of Wisniewski (1931), Siegel (1945) and Gini (1937) who extended Fisher's two-factor formula to three or more factors. Siegel suggested an interesting application of the three-dimensional index, viz., the treatment of changes in wages as a result of changes in labour cost per unit of output, man-hour productivity and man-hours worked. Wisniewski gave indexes of price, area harvested, yield rate and crop value, where crop value was taken as the product of price (per bushel), area harvested (in acres) and yield rate (bushels per acre). Gini considered an example in anthropometry, i.e., in an analysis of variations in the volume of the chest in terms of the variations in antero posterior diameter, transverse diameter of the thorax and sternal height. The generalised formula which they used appear to be much more complicated in form than those suggested in (6.1) based on analogy of 2^m -factorial experiments.

In the case of three or more factors, however, we cannot generalise the result of (3.4); for instance, for three factors

$$V_{111} - V_{000} = P + Q + R + PQR \quad \dots \quad (6.4)$$

i.e., the value change is not equal to sum of the main components unless it is known *a priori* that the second order interaction (PQR) is negligible. Similar difficulties arise in the case of four or more factors. Those difficulties could perhaps be got over and exactly similar formulae worked out by assuming that higher order interactions (that is, interactions higher than three factor interactions) are in practice not very important and could therefore, be ignored. Moreover, in economic applications such high order interactions do not appear to have any obvious physical interpretation.

Assistance of N. S. Iyenger is gratefully acknowledged. His suggestions have been very helpful.

REFERENCES

- BANERJEE, K. S. (1956): Simplification of the derivation of Wald's formula for the cost of living index. *Econometrica*, 24, No. 3, 296-298.
- FRISCH, R. (1936): Annual survey of general economic theory: The problem of index numbers. *Econometrica*, 4, No. 1, 1-38.
- GINI, C. (1937): Methods of eliminating the influence of several groups of factors. *Econometrica*, 5, No. 1, 56-73.

SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS : SERIES A

- SIEGEL, I. H. (1945): The generalised 'ideal' index number formula. *J. Amer. Stat. Ass.*, **40**, No. 232, 520.
- STAEBLE, H. (1935): A development of the economic theory of price index numbers, *The Review of Economic Studies*, 163-188.
- STUVEL, G. (1957): A new index number formula. *Econometrica*, **25**, No. 1, 123-131.
- WALD, A. (1939): A new formula for the index of cost of living. *Econometrica*, **7**, No. 4, 319-331.
- WISNIEWSKI, J. K. (1931): Extension of Fisher's formula number 353 to three or more variables. *J. Amer. Stat. Ass.*, **26**, No. 173, 62.

Paper received : September, 1959.

Revised : October, 1960.

GENERATION OF RANDOM PERMUTATIONS OF GIVEN NUMBER OF ELEMENTS USING RANDOM SAMPLING NUMBERS

By C. RADHAKRISHNA RAO

Indian Statistical Institute

SUMMARY. A general method is given for generating random permutations of integers using a table of random sampling numbers and without wasting the random numbers read. This is more convenient in practice, specially when random permutations of large numbers of elements are needed. It is suggested that even for permutations of small numbers, the method offers greater scope than consulting a table of a limited number of random permutations.

INTRODUCTION

Fisher and Yates (1957) in the introduction to their tables give several methods of obtaining random permutations of integers 1 to n . Some of these methods require random permutations of subsets of integers to build up the whole permutation. A direct method for obtaining a random permutation of integers 1 to n , (or 0 to $n-1$) is to choose k columns in a table of random digits, where $10^k \geq (n-1)$, and write down the numbers 0 to $(n-1)$ in the order they occur, omitting all those above $(n-1)$ and ignoring repetitions. Thus for obtaining a random permutation of numbers 0 to 24 we would take two columns; and the process would naturally reject on an average 75 per cent of the random numbers read. Further, if n is large there is the difficulty of remembering the integers that have not occurred up to any stage while reading the random numbers.

In this note, I give a general method of generating random permutations of integers 1 to n for any n , which does not waste any random number read, and which can be conveniently used even for large n .

ONE-WAY CLASSIFICATION METHOD

Choose a column in a table of random digits and read the digits from any starting point. Take the classes defined by the digits 0 to 9 and write 1 in the class of the first random number, 2 in that of the second number, 3 in that of the third number and so on up to n . For instance, if $n = 7$ and the random numbers read are 5, 3, 9, 1, 2, 3, 1 the integers 1 to 7 are recorded as shown in the first block of Table 1.

TABLE 1. ONE-WAY CLASSIFICATION METHOD

classes defined by digits									
0	1	2	3	4	5	6	7	8	9
	4	5	2		1				3
	7		6						
.....									
		4		7					
.....									
6									2

To obtain a random permutation of integers 1 to 7, first write down the groups of integers in the different classes one after the other starting from the '0' class, which in the present example leads to

(4, 7), 5, (2, 6), 1, 3.

Then randomly permute the subsets within the brackets and remove the brackets. The problem of obtaining random permutations for large numbers is thus reduced to that of obtaining for subsets with expected number in each subset being equal to $1/10$ of the original number. The same process may be used for obtaining permutations of integers within subsets as shown in the last two blocks of Table 1, unless there is a ready way of obtaining these permutations such as referring to tables of random permutations. To permute the subsets continue to read the random numbers. Let the next digits be 2, 4, 9, 1, 8, 0.... Consider the subset (4, 7); 4 goes in class 2 and 7 in class 4, which completes the permutation. Similarly a permutation of (2, 6) is derived in the last block of Table 1, and the complete permutation is obtained as

4, 7, 5, 6, 2, 1, 3.

If permutations for large values of n are needed, several classes will have multiple entries, needing one or more repetitions of the process for each subset. This can be avoided to a large extent by following the two-way classification method and, if convenient, by adopting the built-in procedure of simultaneously permuting within subsets while recording the integers, as detailed in the next section.

TWO-WAY CLASSIFICATION METHOD

Choose two columns in a table of random digits. Each random number would now consist of two digits which define a cell in a 10×10 two-way table as shown in Table 2.

TABLE 2. TWO-WAY CLASSIFICATION METHOD

		second digit									
		0	1	2	3	4	5	6	7	8	9
first digit	0		9				10				
	1								2		
	2				6		12				
	3		1,5						7		11
	4						4,16				
	5				17						
	6		8							15	
	7					13		18			
	8	3									
	9										14

Let a random permutation of numbers 1 to 18 be needed. Write 1 in the cell of the first random number, 2 in the cell of the second number, ... and so on up to 18. For instance, corresponding to the first five random numbers 31, 17, 81, 45, 31 the entries 1 to 5 would be as shown in Table 2. Proceeding further in the same manner, the first-18 random numbers are used

GENERATION OF RANDOM PERMUTATIONS USING RANDOM SAMPLING NUMBERS

to post all the integers 1 to 18 in the cells. In two cells there are two entries, which could have been avoided by omitting a number if its cell is already filled. Otherwise the integers within subsets (1, 5) and (4, 16) have to be randomly permuted first. Suppose (5, 1) and (4, 16) are the permutations arrived at by some process. The entire permutation could be read out in any convenient manner, say from left to right row by row, the order of integers in cells with multiple entries being determined by independent permutations. In the present example the permutation is

9, 10, 2, 6, 12, 5, 1, 7, 11, 4, 16, 17, 8, 15, 13, 18, 3, 14.

The problem of multiple entries, due to repetition of random numbers, is not serious since the chance of having more than two or three entries in a cell is small if n is not very large, and it is easy to obtain a random permutation of two or three integers.

It is also possible for the random permutation of multiple entries to be built into the process of entering the numbers in the cells as follows. For instance, when 5 has to be placed in the cell (31), which already has 1, read the digit in the same row containing 31 in an adjacent column of the random number table. If this digit is even, write 5 (the second number to be placed in the cell) before 1 (the first number), and if odd, write 5 after 1. Random permutation of 1, 5 is already secured by this operation, and similarly the order of integers in (4, 16) could have been settled while posting 16 and the whole random permutation is obtained straight away from the two-way table. If a random number repeats for a third time, first read the digit in the same row in an adjacent column. If this is 0, pass on to the next column, and so on, till a non-zero digit is obtained from the row. The number to be written in the cell is put in the first position if a non-zero digit read from an adjacent column is $\equiv 0 \pmod{3}$, in the second position if $\equiv 1 \pmod{3}$, and in the third position if $\equiv 2 \pmod{3}$. A similar device could be adopted in case a random number repeats for a fourth time, and so on. As stated earlier, the permutation of multiple entries in the cells may be undertaken at the final stage, if that is more convenient.

For larger values of n , to avoid multiple entries, one could use three columns of a random number table and consider a triple coordinate system to identify 1000 cells of a three-way table. The best way of representing would be a triple classification table involving 10 two-way sub-tables, each sub-table corresponding to a given value of the third digit. The procedure outlined before can then be followed.

It is important to note that for any value of n , one can use any arbitrary number k of columns of the random digits and proceed to distribute the integers in the cells of a k -way table. But it may be convenient to use one-way classification method for $n \leq 10$, two-way classification method for $n \leq 100$ and so on.

In this connection, the reader is referred to two papers by Lahiri (1951) and Matthai (1953) on choosing, from a list of sampling units, with unequal and equal probabilities respectively, a random sample with the help of random sampling numbers.

REFERENCES

- FISHER, R. A. and YATES, F. (1957): *Statistical Tables for Biological, Agricultural and Medical Research*, 5th ed. Edinburgh, Oliver and Boyd.
- LAHIRI, D. B. (1951): A method for sample selection providing unbiased ratio estimates. *Bull. Inst. Int. Stat.*, **33**, 2, 133-140.
- MATTHAI, ABRAHAM (1953): On selecting random numbers for large-scale sampling. *Sankhyā*, **13**, 257-260.

Paper received : July, 1961.

CORRIGENDA

Tables of Random Normal Deviates : By J. M. Sengupta and N. Bhattacharya, *Sankhyā*, **20**, 249-286.

- (1) Plate 2, col. 8, line 28 : should read + 2.260 and *not* + 2.268.
- (2) Plate 2, col. 8, line 29 : should read + 1.195 and *not* + 1.190.
- (3) Plate 2, col. 8, line 30 : should read + 1.298 and *not* 1.295.

On Some Classification Problems-1 : By S. John, *Sankhyā*, **22**, 301-308.

Expression (2.11) : First factor is $f(w_1^{(1)}, w_1^{(2)})$.

Equation (3.8) : The numerator of the last factor of the summand should be

$$\Gamma\left(\frac{p-1}{2}\right) \Gamma\left(\frac{2k+s+2}{2}\right)$$

On Some Classification Statistics: By S. John, *Sankhyā*, **22**, 309-316.

Equation (3.2) : Reverse the inequality sign.

Equation (5.3) : Change b_2^2 to $2b_2^2$.

ACKNOWLEDGEMENT

Thanks are due to the following for their kind help in the editing of the papers published in this issue.

Dr. D. Basu

Shri R. N. Bhattacharyya

Shri N. S. Iyengar

Dr. Sujit Kumar Mitra

Shri K. R. Parthasarathy

Dr. C. R. Rao

Shri S. S. Shrikhande

Shri R. Ranga Rao

Dr. J. Roy

Shri S. R. S. Varadhan

Editor,

Sankhyā : The Indian Journal of Statistics.

BIOMETRIKA

Volume 48, Parts 3 and 4

December, 1961

CONTENTS

The two sample capture-recapture census when tagging and sampling are stratified.....	Darroch, J. N.
The fiducial method and invariance.....	Fraser, D. A. S.
Bayesian sequential analysis.....	Wetherill, G. B.
Some properties of the Spearman estimator in bioassay.....	Brown, Byron Wm. (Jr.)
Inconsistencies in a schedule of paired comparisons.....	Slater, Patrick
Some circular coverage problems.....	Morgenthaler, George W.
Ordered tests in the analysis of variance.....	Bartholomew, D. J.
Quota fulfilment using unrestricted random sampling.....	Young, D. H.
Large-sample estimation of parameters for moving-average models.....	Walker, A. M.
Linear and non-linear multiple comparisons in logit analysis.....	Reiersol, Olav
Estimation of the normal population parameters given a type I censored sample.....	Saw, J. G.
Some tests for outliers.....	Quesenberry, C. P. and David, H. A.
The ergodic behaviour of random walks.....	Kingman, J. F. C.
The goodness-of-fit of a single (non-isotropic) hypothetical principal component.....	Kshirsagar, A. M.
On the evaluation of the probability integral of the multivariate t -distribution.....	John, S.
Computing the distribution of quadratic forms in normal variables.....	Inhof, J. P.
Comparative efficiencies of methods of estimating parameters in linear autoregressive schemes.....	Chanda, K. C.
Tables of the Freeman-Tukey transformations for the binomial and Poisson distributions.....	Mosteller, Frederick and Youtz, Cleo
Miscellanea: Contributions by J. D. Biggers; E. J. Burr and Gwenda Cane; C. F. Crouse; P. D. Finch; Gerald J. Glasser and Robert F. Winter; D. Hogben, R. S. Pinkham and M. B. Wilk; B. K. Kale; J. G. Saw; A. Zinger.	
Reviews	
Other Books Received	
Corrigenda	

The subscription, payable in advance, is 54/- (or \$8.00), per volume (including postage). Cheques should be made payable to *Biometrika*, crossed "A/c *Biometrika Trust*" and sent to the Secretary, *Biometrika* Office, Department of Statistics, University College, London, W. C. 1. All foreign cheques must be drawn on a Bank having a London agency.

Issued by THE BIOMETRIKA OFFICE, University College, London.

JOURNAL OF THE ROYAL STATISTICAL SOCIETY

Series B (Methodological)

40s each Part. Annual Subscription £3.2s (US \$9). Post free

CONTENTS OF VOL. 23, PART 2, 1961

A test of homogeneity of means under restricted alternatives (With Discussion).....	D. J. Bartholomew
Optimal programmes for varietal selection (With Discussion).....	R. N. Curnow
Some stochastic processes with absorbing barriers.....	T. L. Saaty
Some sequential tests using range.....	W. G. Gilchrist
First emptiness in a finite dam.....	B. Weesakul
The construction of optimal designs for the one-way classification analysis of variance.....	J. Aitchison
Some simple wear-dependent renewal processes.....	A. Mercer
An exact distribution of the Fisher-Behrens-Welch statistic for testing the difference between the means of two normal populations with unknown variances.....	W. D. Ray and A. E. N. T. Pitman
Closed queueing systems: A Generalisation of the machine interference model.....	F. Benson and G. Gregory
Testing for a jump in the spectral function.....	E. J. Hannan
Moments of a truncated bivariate normal distribution.....	S. Rosenbaum
On quadratic estimates of the interclass variance for unbalanced designs.....	R. R. Read
Mixtures of geometric distributions.....	H. E. Daniels
Prediction by exponentially weighted moving averages and related methods.....	D. R. Cox
The randomization bases of the problem of the amalgamation of weighted means.....	D. E. Barton and C. L. Mallows
Estimation from quantiles in destructive testing.....	H. P. Stout and F. Stern
A note on maximum-likelihood in the case of dependent random variables.....	S. D. Silvey
On some properties of the asymptotic variance of the sample quantiles and mid-ranges.....	P. K. Sen
Similarities between likelihoods and associated distributions <i>a posteriori</i>	D. A. Sprott
A method of maximum likelihood estimation.....	F. S. G. Richards
The intervals between regular events displaced in time by independent random deviations of large dispersion.....	T. Lewis
An approach to the scheduling of jobs on machines.....	E. Page

Obtainable through any bookseller or direct from the

ROYAL STATISTICAL SOCIETY

21 Bentinck Street, London, W. 1.

THE ANNALS OF MATHEMATICAL STATISTICS

(Founded by H. C. Carver)

The Official Journal of the Institute of Mathematical Statistics

CONTENTS OF VOLUME 32, NO. 4, DECEMBER, 1961

An Approach to Time Series Analysis.....	Emanuel Parzen
Some Model I Problems of Selection.....	E. L. Lehmann
Bayes' Rules for a Common Multiple Comparisons Problem and Related Student- <i>t</i> Problems.....	David B. Duncan
The Use of Least Favorable Distributions in Testing Composite Hypotheses.....	H. E. Reinhardt
Asymptotic Efficiency in Polynomial Estimation.....	Paul G. Hoel
Maximum Likelihood Estimation of a Linear Functional Relationship.....	C. Villegas
Sequential χ^2 - and T^2 -Tests.....	Edward J. Jackson and Ralph A. Bradley
Estimating the Parameters of Negative Exponential Populations from One or Two Order Statistics.....	H. Leon Harter
On the Two Sample Problem: A Heuristic Method for Constructing Tests.....	V. P. Godambe
A Nonparametric Test for the Problem of Several Samples.....	V. P. Bhapkar
Distribution of the Anderson-Darling Statistic.....	Peter A. W. Lewis
Errors in Discrimination.....	S. John
On the Monotonic Character of the Power Functions of Two Multivariate Tests.....	S. N. Roy and W. F. Mikhall
The Moments of Elementary Symmetric Functions of the Roots of a Matrix in Multivariate Analysis.....	Tito A. Mijares
Variance Components in the Unbalanced 2-Way Nested Classification.....	S. R. Searle
Some Main-Effect Plans and Orthogonal Arrays of Strength Two.....	Sidney Addelman and Oscar Kempthorne
On A Geometrical Method of Construction of Partially Balanced Designs with Two Associate Classes.....	Esther Seiden
On Some Methods of Construction of Partially Balanced Arrays.....	I. M. Chakravarti
Some Further Designs of Type 0: PP.....	G. H. Freeman
Sufficiency in the Undominated Case.....	D. L. Burkholder
On a Special Class of Recurrent Events.....	M. P. Schützenberger
Maximum Likelihood Characterization of Distributions.....	Henry Teicher
On the Distribution of First Significant Digits.....	Roger S. Pinkham
Markov Renewal Processes: Definitions and Preliminary Properties.....	Ronald Pyke
Markov Renewal Processes with Finitely Many States.....	Ronald Pyke
A Convexity Property in the Theory of Random Variables Defined on a Finite Markov Chain.....	H. D. Miller
Limit Distribution in the Theory of Counters.....	G. Sankaranarayanan
The Transient Behavior of a Single Server Queuing Process with Recurrent Input and Gamma Service Time.....	Lajos Takács
Efficient Estimation of a Regression Parameter for Certain Second Order Processes.....	Charlotte T. Striebel
Two Similar Queues in Parallel.....	J. F. C. Kingman
Queues with Batch Departures I.....	F. G. Foster and K. M. Nyunt
NOTES	
On the Chapman-Kolmogorov Equation.....	Jack Karush
A Generalization of a Theorem of Balakrishnan.....	N. Donald Ylvisaker
The Opinion Pool.....	M. Ston
Correction Notes	
Abstracts of Papers	
News and Notices	
Report of the President for 1961	
Report of the Editor for 1961	
Publications Received	

Subscription Rate: \$ 10.00 per year

Address orders for subscriptions and back numbers to the

Secretary, Institute of Mathematical Statistics,

Department of Statistics, University of

North Carolina, Chapel Hill, North Carolina.

ECONOMETRICA

Vol. 29, No. 3

July, 1961

CONTENTS

A Model of Japanese Economic Growth, 1878-1937.....	L. R. Klein
Manufacturers' Inventories, Sales Expectations, and the Acceleration Principle.....	Michael C. Lovell
Rational Expectations and the Theory of Price Movements.....	John F. Muth
Efficient Allocation of a Multi-Purpose Sample.....	Leslie Kish
The Place of Least Squares in Econometrics.....	Frederick V. Waugh
The Interpretation of Cross Section Estimates in a Dynamic Model.....	Yehuda Grunfeld
A Reconsideration of Domar's Theory of Economic Growth.....	Ragnar Frisch
The Maximum Likelihood Estimation of Economic Relationships with Auto-Regressive Residuals.....	Denis Sargan
The Statistical Discrepancy in the National Accounts Revisited.....	Peter E. de Janosi
The Estimation of Distributed Lags: A Comment.....	E. Malinvaud
Report of the Stanford Meeting	
Report of the St. Louis Meeting	
Book Reviews	
Announcements and Notes	

Subscription price for nonmembers : \$ 12.00 per year in the United States and Canada;

\$ 10.00 per year in all other Countries. Single copies \$ 3.00

Graduate Students recommended by a faculty member : \$ 4.00 per year.

Subscriptions should be sent to

**The Secretary
ECONOMETRIC SOCIETY**

Box 1264, Yale Station, New Haven, Connecticut, U.S.A.

SANKHYĀ

THE INDIAN JOURNAL OF STATISTICS

Edited by : P. C. MAHALANOBIS

SERIES A, VOL. 23

NOVEMBER 1961

PART 4

THE TRANSFORMATION OF A DISTRIBUTION UNDER SELECTION

By D. J. FINNEY

University of Aberdeen, and Agricultural Research Council Unit of Statistics

SUMMARY. For a general bivariate frequency distribution, specified by its cumulants, results have been obtained pertaining to the effect on the distribution of the first variate produced by selection in respect of the second variate. The most general rule of selection studied is that which defines the probability of selection as a function of the second variate, and the moment generating function for the first variate after selection is stated as equation (5.5). This concise formula appears to be of little direct use for calculation, though an interesting expansion in series is possible when the initial distribution is bivariate normal, equations (4.10), (4.16), and (4.19).

The main task of the latter part of the paper has been to obtain explicit expressions, for the first four moments of the first variate after selection, in the case of prime importance, namely "cut-off selection" in which all values above a certain limit are taken and all others discarded. The moments have been given in two forms, in terms of the cut-off point and in terms of the proportion selected and the corresponding normal deviate.

1. INTRODUCTION

In an earlier paper (Finney, 1956), I have examined the nature of the frequency distribution obtained when a variate that can be measured only with the inclusion of a normally distributed error is subjected to a "cut-off" selection based upon measured values. That is to say, each individual measured is rejected unless its *measurement* exceeds an arbitrary quantity, and the interest then lies in the frequency distribution of *true values* among those selected. The arbitrary cut-off point was there defined in terms of the proportion of the whole population to be selected, and expansions in series for the first four moments of the true values selected were obtained. This is a situation that arises in many practical circumstances. For example, in the selection of crop or animal strains or varieties (Cochran, 1951; Finney, 1958; Robertson, 1957), the variate whose values are to be increased by selection will be the yielding capacity of a genotype, but decisions must be made on the basis of phenotypic performance.

Again, in the selection of children for secondary education, innate ability may be the character of real interest, but the evidence on which decisions are to be taken will consist of scores in examinations and tests of intelligence and skill. The assumption that the measured value deviates from the true by a normally distributed error is likely to be adequate for the theoretical study of such problems.

Results of greater generality may sometimes be useful. With this in view, not only has the restriction to a normal error distribution been removed but also a less rigid process of selection than a simple cut-off has been studied. Certain formal results of greater generality than before have been derived.

2. SPECIFICATION OF INITIAL DISTRIBUTIONS

For the concise expression of general results, the notation of the previous paper needs to be substantially modified. Suppose x, y are two variates with probability density function $p(x, y)$ for which all cumulants, κ_{ij} , exist. Then, with the convention that all integrals are over the interval $(-\infty, \infty)$ unless the contrary is explicit*,

$$\begin{aligned} \int \int \exp(\theta x + \phi y) p(x, y) dx dy &= M_{x,y}(\theta, \phi) \\ &= \exp \left(\sum \sum \frac{\kappa_{ij}}{i!j!} \theta^i \phi^j \right) \end{aligned} \quad \dots (2.1)$$

defines the relation between the p.d.f., the moment generating function, and the cumulants. (Throughout the paper, θ and ϕ are used in preference to $i\theta$ and $i\phi$ for simplicity of writing, as this makes no essential change in the mathematical development). Here x is to be regarded as the variate in which an investigator is primarily interested, but selection is to be based on the values of y and the effect of this selection on the distribution of x then studied. This situation for which this theory was developed in 1956 was that of x representing the genotypic value of some character for an individual, and y representing the phenotypic expression of this character; y can be measured directly, but individual values of x cannot be measured and the nature of the distribution of x after selection can only be inferred from that of y .

One important class of problems, to which the genetic situation just mentioned belongs, has

$$y = x + w, \quad \dots (2.2)$$

where w is an error of observation or measurement from which x cannot be freed. The 1956 paper was restricted to this model, with w normally distributed and independent of x . The theory that follows is applicable whatever the distribution of w and however x and w may be related as long as all cumulants exist. If $M_{x,w}^*(\theta, \phi)$ is the m.g.f. of the joint distribution of x and w ,

$$\begin{aligned} M_{x,y}(\theta, \phi) &= \int \int \exp[(\theta + \phi)x + \phi w] p(x, x+w) dx dw \\ &= M_{x,w}^*(\theta + \phi, \phi). \end{aligned} \quad \dots (2.3)$$

*If the ranges permitted to the variates are finite, the limits of integration will be appropriately modified and similar changes will affect definitions such as equation (3.1).

THE TRANSFORMATION OF A DISTRIBUTION UNDER SELECTION

If now κ_{ij}^* denotes the cumulants of x and w ,

$$\sum_i \sum_j \frac{\kappa_{ij}}{i!j!} \theta^i \phi^j = \sum_i \sum_j \frac{\kappa_{ij}^*}{i!j!} (\theta + \phi)^i \phi^j \quad \dots (2.4)$$

and equation of coefficients enables the κ_{ij} to be expressed in terms of the κ_{ij}^* . Thereafter formulae appropriate to the model indicated by (2.2) are based upon the general results of this paper.

Without real loss of generality, means may be taken as origins, so that

$$\kappa_{10} = 0, \quad \kappa_{01} = 0. \quad \dots (2.5)$$

A further convenient symbolism is

$$\kappa_{20} = \sigma^2, \quad \kappa_{11} = \rho\sigma\omega, \quad \kappa_{02} = \omega^2 \quad \dots (2.6)$$

σ^2 and ω^2 being the variances of x and y and ρ their correlation coefficient. A generalization of a result stated by Cornish and Fisher (1937) enables $p(x, y)$ to be expressed formally in terms of the cumulants and the differential operators

$$D_x = \frac{\partial}{\partial x} \quad D_y = \frac{\partial}{\partial y}; \quad \dots (2.7)$$

this is
$$p(x, y) = \exp \left[-\frac{1}{2}(\sigma^2 D_x^2 + 2\rho\sigma\omega D_x D_y + \omega^2 D_y^2) \right] M_{x,y}(-D_x, -D_y) \\ \cdot \frac{1}{2\pi\sigma\omega(1-\rho^2)^{\frac{1}{2}}} \exp \left[-\frac{1}{2(1-\rho^2)} \left(\frac{x^2}{\sigma^2} - \frac{2\rho xy}{\sigma\omega} + \frac{y^2}{\omega^2} \right) \right]. \quad \dots (2.8)$$

Also, the marginal p.d.f. for y alone is

$$p_1(y) = \exp \left(-\frac{1}{2}\omega^2 D_y^2 \right) M_{x,y}(0, -D_y) \frac{1}{\omega\sqrt{2\pi}} e^{-\frac{y^2}{2\omega^2}}. \quad \dots (2.9)$$

The m.g.f. for x -array corresponding to any specified y , $M_{x|y}(\theta)$, can now be concisely expressed by means of Bartlett's theorem (1938a):

$$M_{x|y}(\theta) \cdot p_1(y) = \frac{M_{x,y}(\theta, -D_y)}{M_{x,y}(0, -D_y)} p_1(y) \\ = \exp \left(-\frac{1}{2}\omega^2 D_y^2 \right) M_{x,y}(\theta, -D_y) p_1(y). \quad \dots (2.10)$$

The variates may usefully be standardized, so as to have unit standard deviations initially, by writing

$$y = \omega u, \quad \dots (2.11)$$

$$D_y = \omega^{-1} \frac{\partial}{\partial u} = \omega^{-1} D, \quad \text{say,} \quad \dots (2.12)$$

and

$$\kappa_{ij} = \gamma_{ij} \sigma^i \omega^j. \quad \dots (2.13)$$

In particular, of course,

$$\gamma_{10} = \gamma_{01} = 0; \quad \text{and} \quad \gamma_{20} = \gamma_{02} = 1, \quad \gamma_{11} = \rho.$$

Then

$$M_{x|y}(\theta) p_1(y) dy = \exp \left(-\frac{1}{2} D^2 \right) \exp \left[\sum_i \sum_j \frac{\gamma_{ij}}{i!j!} (\sigma\theta)^i (-D)^j \right] \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du. \quad (2.14)$$

3. THE SELECTOR FUNCTION

A general process of selection operating on y (either for the whole population or for a random sample thereof) may be represented by the probability that an individual with a particular value of y is selected. This quantity, hereafter referred to simply as the *selector*, will be written as a function $h\left(\frac{y}{\omega}\right)$ such that

$$h\left(\frac{y}{\omega}\right) \geq 0 \quad (-\infty < y < \infty) \quad \dots (3.1)$$

and, without loss of generality,

$$\max \left[h\left(\frac{y}{\omega}\right) \right] = 1, \quad \dots (3.2)$$

because only the proportions in which different values of y are selected affect results. The selector is not itself a p.d.f. Its definition means that the p.d.f. of Y , which symbol will be used to denote y in the selected portion of the population, is $p_2(Y)$, where

$$P = \int h\left(\frac{y}{\omega}\right) p_1(y) dy \quad \dots (3.3)$$

is the proportion of the whole population (or the expected proportion of the measured random sample) that is selected and

$$p_2(Y) = P^{-1} h\left(\frac{Y}{\omega}\right) p_1(Y). \quad \dots (3.4)$$

The problem under discussion is that of finding properties of the frequency distribution of X , the value of x in the selected portion of the population.

The concept of a selector includes the cut-off selection of the previous paper, for which

$$h\left(\frac{y}{\omega}\right) = \begin{cases} 0 & (y < \eta) \\ 1 & (y \geq \eta) \end{cases} \quad \dots (3.5)$$

and η is defined so as to ensure that a specified proportion of the y -distribution is selected. A selector corresponding to a double cut-off may also be of interest; it is

$$h\left(\frac{y}{\omega}\right) = \begin{cases} 1 & (\eta_1 < y < \eta_2) \\ 0 & (y \leq \eta_1 \text{ or } y \geq \eta_2) \end{cases} \quad \dots (3.6)$$

if the central portion is to be that selected, or this with all inequalities reversed if the two tails are to be selected. The interpretation of general results so as to apply to selectors such as (3.5) and (3.6) must be undertaken with care because of discontinuities.

4. THE NORMAL DISTRIBUTION

The simplest version of the general problem is of intrinsic importance and also of interest as indicating the type of result that is wanted. It is however scarcely typical of the complexity that will be encountered later. Suppose that $p(x, y)$ is a bivariate normal p.d.f., so that

$$\kappa_{ij} = 0 \quad \text{for } i+j > 2. \quad \dots (4.1)$$

THE TRANSFORMATION OF A DISTRIBUTION UNDER SELECTION

The m.g.f. of Y , from (3.4), is

$$\begin{aligned} M_Y(\phi) &= \frac{1}{P\omega\sqrt{2\pi}} \int h\left(\frac{Y}{\omega}\right) e^{-\frac{Y^2}{2\omega^2} + \phi Y} dY \\ &= \frac{1}{P\sqrt{2\pi}} e^{\frac{1}{2}\omega^2\phi^2} \int h(u + \omega\phi) e^{-\frac{1}{2}u^2} du \\ &= \frac{e^{\frac{1}{2}\omega^2\phi^2} E[h(u + \omega\phi)]}{E[h(u)]}, \end{aligned} \quad \dots (4.2)$$

where the expectation function $E[\]$ is always to be understood as taken with respect to the unit normal p.d.f.: by definition, for any function $g(u)$,

$$E[g(u)] = \frac{1}{\sqrt{2\pi}} \int g(u) e^{-\frac{1}{2}u^2} du. \quad \dots (4.3)$$

Now the joint distribution of x, y can be regarded as generated by a normal distribution of y (mean 0 and variance ω^2) and a linear regression of x on y such that x has a variance $\sigma^2(1-\rho^2)$ about its expectation βy , where

$$\beta = \rho\sigma/\omega. \quad \dots (4.4)$$

This regression of x on y is clearly unaffected by the selection. Hence the m.g.f. of X , the value taken by the variate X after selection, is

$$M_X(\theta) = \frac{e^{\frac{1}{2}\sigma^2\theta^2} E[h(u + \beta\omega\theta)]}{E[h(u)]}. \quad \dots (4.5)$$

Write

$$h_i(u) = D^i h(u) \quad \dots (4.6)$$

and

$$E_i = E[h_i(u)]. \quad \dots (4.7)$$

Also define an operator Δ having the property

$$\Delta E_i = E_{i+1}, \quad \dots (4.8)$$

and such that when applied to any function of the E_i it behaves as though the suffixes are exponents and it is a differential operator; for example

$$\left. \begin{aligned} \Delta(E_i E_j) &= E_{i+1} E_j + E_i E_{j+1} \\ \Delta(E_i^2) &= 2E_i E_{i+1} \\ \Delta\left(\frac{1}{E_i}\right) &= -\frac{E_{i+1}}{E_i^2} \\ \Delta(\log E_i) &= \frac{E_{i+1}}{E_i} \end{aligned} \right\} \quad \dots (4.9)$$

Then the cumulant function for X can readily be expanded as a power series in θ :

$$\begin{aligned}\log M_X(\theta) &= \frac{1}{2}\sigma^2\theta^2 - \log E_0 + \log\{E[h(u + \beta\omega\theta)]\} \\ &= \frac{1}{2}\sigma^2\theta^2 + \sum_{i=1}^{\infty} (\beta\omega)^i \Delta^i(\log E_0) \frac{\theta^i}{i!}.\end{aligned}\quad \dots \quad (4.10)$$

Hence the cumulants of x for the distribution of selected individuals are

$$\left. \begin{aligned}K_1 &= \frac{\beta\omega E_1}{E_0}, \\ K_2 &= \sigma^2 + \frac{(\beta\omega^2)(E_2 E_0 - E_1^2)}{E_0^2}, \\ K_i &= (\beta\omega)^i \Delta^i(\log E_0) \quad \text{for } i > 2.\end{aligned}\right\} \quad \dots \quad (4.11)$$

An alternative expression for E_i deserves noting. If

$$Z(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}, \quad \dots \quad (4.12)$$

$$\text{then} \quad E_i = \int h_i(u) Z(u) du = \int h(u) (-D)^i Z(u) du, \quad \dots \quad (4.13)$$

this following from i successive integrations by parts.

If now the selector (3.5) is considered, evidently

$$P = \int_U^{\infty} Z(u) du, \quad \dots \quad (4.14)$$

where

$$U = \eta/\omega. \quad \dots \quad (4.15)$$

By the use of (4.13), the discontinuity of the selector at $u = U$ is seen to lead to

$$E_0 = P, \quad E_1 = Z(U), \quad \Delta = - \frac{d}{dU}.$$

Therefore, (4.11) can be written

$$\left. \begin{aligned}K_2 &= \sigma^2 + \beta^2\omega^2 \frac{d^2}{dU^2} (\log P), \\ K_i &= (-\beta\omega)^i \frac{d^i}{dU^i} (\log P) \quad \text{for } i = 1, 3, 4, \dots\end{aligned}\right\} \quad \dots \quad (4.16)$$

This agrees exactly with equations (12) of the previous paper, although the definition of β is somewhat wider. The model previously used was based upon equation (2.2), a special case of that now under discussion; for it

$$\kappa_{11} = \sigma^2 \text{ so that } \rho = \sigma/\omega, \text{ and } \beta = \sigma^2/\omega^2.$$

THE TRANSFORMATION OF A DISTRIBUTION UNDER SELECTION

The selector (3.6), which might be used if the aim of selection were to homogenize the population by rejection of extreme values rather than to increase the mean, gives similar formulae. If

$$U_1 = \eta_1/\omega, \quad U_2 = \eta_2/\omega, \quad \dots \quad (4.17)$$

then
$$P = \int_{U_1}^{U_2} Z(u) du. \quad \dots \quad (4.18)$$

Hence

$$E_0 = P, \quad E_1 = Z(U_1) - Z(U_2),$$

and

$$\Delta = \frac{\partial}{\partial U_2} - \frac{\partial}{\partial U_1}.$$

Consequently

$$\left. \begin{aligned} K_2 &= \sigma^2 + \beta^2 \omega^2 \left(\frac{\partial}{\partial U_2} - \frac{\partial}{\partial U_1} \right)^2 \log P \\ K_i &= (\beta \omega)^i \left(\frac{\partial}{\partial U_2} - \frac{\partial}{\partial U_1} \right)^i \log P \\ &\quad \text{for } i = 1, 3, 4, \dots \end{aligned} \right\} \quad \dots \quad (4.19)$$

5. GENERAL THEORY

When $p(x, y)$ is completely general in form, the method of approach to the properties of the distribution of X must be changed. Pearson (1912) obtained what appeared to be general results (at least for means and variances) with a multivariate population, but his analysis contained an implicit assumption of linearities of regressions. Bartlett (1938b) and Lawley (1943), while showing that this did not necessitate normality of distributions, demonstrated that some rather special conditions must be satisfied. Even if the regression of y on x is linear, the methods of Section 4 will break down unless linearity of the regression of x on y permits moments of X to be derived from moments of Y . For the model in equation (2.2), Lindley (1947) proved the necessary and sufficient condition that x has a linear regression on y to be that the cumulant generating function of x is a multiple of that of w , and Curnow (1960) has further studied this very special situation. For the general problem of this paper, a different approach seems essential.

Equation (2.10) gives the m.g.f. of x for any y . Clearly $M_{x|Y}(\theta)$ and $M_{x|y}(\theta)$ are identical wherever the former is defined, that is to say for any y that has a non-zero probability of selection. Hence $M_X(\theta)$ can be obtained by integration of $M_{x|y}(\theta)$ over the frequency distribution of Y :

$$\begin{aligned} M_X(\theta) &= \int M_{x|Y}(\theta) p_2(Y) dY \\ &= P^{-1} \int h(u) e^{-\frac{1}{2} D^2} \exp \left[\sum \sum \frac{\gamma_{ij}}{i! j!} (\sigma \theta)^i (-D)^j \right] Z(u) du \quad \dots \quad (5.1) \end{aligned}$$

by (3.4) and (2.14). Moreover, from (2.9) and (3.3) or from the terms independent of θ in (5.1),

$$P = \int h(u) e^{-\frac{1}{2}D^2} \exp \left[\sum \frac{\gamma_{0j}}{j!} (-D)^j \right] Z(u) du. \quad \dots (5.2)$$

Write
$$f(\theta, D) = \exp \sum_3^\infty \sum_3^\infty \frac{\gamma_{ij}}{i!j!} (\sigma\theta)^i (-D)^j \quad \dots (5.3)$$

and rearrange the terms involving cumulants of the second order in (5.1), (5.2).

Then

$$M_X(\theta) = \frac{e^{\frac{1}{2}\sigma^2\theta^2} \int h(u) f(\theta, D) Z(u - \rho\sigma\theta) du}{\int h(u) f(0, D) Z(u) du}. \quad \dots (5.4)$$

This can be put into a form more reminiscent of (4.5); integration by parts, as for equation (4.13), leads to

$$M_X(\theta) = \frac{e^{\frac{1}{2}\sigma^2\theta^2} E[f(\theta, -D) h(u + \rho\sigma\theta)]}{E[f(0, -D) h(u)]}. \quad \dots (5.5)$$

When $p(x, y)$ is bivariate normal, $f(\theta, -D)$ is identically unity and, since

$$\rho\sigma = \beta\omega,$$

(5.5) reduces to (4.5). From (5.5), by formal expansion in powers of θ , the cumulant function of X can be expressed in a manner similar to (4.10).

6. GENERAL CUT-OFF SELECTION

Although (5.5) is a concise expression of the most general result, (5.1) is more useful in the development of explicit formulae for moments. The process will be illustrated for simple cut-off selection, again with (3.5) as the selector. The process is the same as in the previous (1956) paper, but results for a general p.d.f. of x, y are now available and the expansions in series have been taken to higher order.

The first step is to expand the operator in (2.14),

$$F(D) = e^{-\frac{1}{2}D^2} \exp \left[\sum \sum \frac{\gamma_{ij}}{i!j!} (\sigma\theta)^i (-D)^j \right], \quad \dots (6.1)$$

in powers of $(\sigma\theta)$ and D . The general term in the expansion is

$$\frac{1}{p!q!r! \dots (i!i'!)^p (j!j'!)^q (k!k'!)^r \dots} \gamma_{ii'}^p \gamma_{jj'}^q \gamma_{kk'}^r \dots (\sigma\theta)^{pi+qj+rk} \dots (-D)^{pi'+qj'+rk'} \dots \quad \dots (6.2)$$

except that any term involving γ_{02} is to be omitted, in order to take account of the factor $e^{-\frac{1}{2}D^2}$. Cornish and Fisher (1937) have suggested that the cumulants of a dis-

THE TRANSFORMATION OF A DISTRIBUTION UNDER SELECTION

tribution will often be in decreasing order of magnitude in respect of a parameter, n , representing the size of a sample. An appropriate supposition, in line with Cornish and Fisher, is that

$$\gamma_{ij} = 0(n^{1-i(i+j)}) \quad \text{for } i+j \geq 2. \quad \dots (6.3)$$

The expansion of $F(D)$ has been arranged in powers of $(\sigma\theta)$; the terms in $(\sigma\theta)$ have been taken to $0(n^{-2})$, and those in $(\sigma\theta)^2$, $(\sigma\theta)^3$, and $(\sigma\theta)^4$ to $0(n^{-3/2})$. This expansion is not written out here, but it is of a pattern similar to that of equation (20) in the previous paper.

Suppose that η has been so chosen as to cut off a proportion P of the population; write

$$U = \eta/\omega, \quad \dots (6.4)$$

and regard P as a function of U . Then (3.3) becomes

$$P(U) = \int_{\eta}^{\infty} p_1(y)dy. \quad \dots (6.5)$$

Define T as the unit normal deviate corresponding to a specified probability $P_N(T)$; i.e.

$$P_N(T) = \int_T^{\infty} Z(u)du, \quad \dots (6.6)$$

with $Z(u)$ as defined in (4.12). A result of Cornish and Fisher (1937), extended by Fisher and Cornish (1960), now enables the value of U corresponding to a particular T to be written as an infinite series; if

$$P(U) = P_N(T), \quad \dots (6.7)$$

then a consequence of (2.9) is that, to $0(n^{-2})$,

$$\begin{aligned} U = T &+ \frac{1}{6} \gamma_{03}(T^2-1) + \frac{1}{24} \gamma_{04}(T^3-3T) - \frac{1}{36} \gamma_{03}^2(2T^3-5T) \\ &+ \frac{1}{120} \gamma_{05}(T^4-6T^2+3) - \frac{1}{24} \gamma_{03}\gamma_{04}(T^4-5T^2+2) \\ &+ \frac{1}{324} \gamma_{03}^3(12T^4-53T^2+17) + \frac{1}{720} \gamma_{06}(T^5-10T^3+15T) \\ &- \frac{1}{180} \gamma_{03}\gamma_{05}(2T^5-17T^3+21T) - \frac{1}{384} \gamma_{04}^2(3T^5-24T^3+29T) \\ &+ \frac{1}{288} \gamma_{03}^2\gamma_{04}(14T^5-103T^3+107T) - \frac{1}{7776} \gamma_{03}^4(252T^5-1688T^3+1511T). \dots (6.8) \end{aligned}$$

Expansion of $Z(U)$ then leads to

$$\begin{aligned}
 Z(U) = Z(T) & \left[1 - \frac{1}{6} \gamma_{03}(T^3 - T) - \frac{1}{24} \gamma_{04}(T^4 - 3T^2) + \frac{1}{72} \gamma_{03}^2(T^6 + T^4 - 7T^2 - 1) \right. \\
 & - \frac{1}{120} \gamma_{05}(T^5 - 6T^3 + 3T) + \frac{1}{144} \gamma_{03} \gamma_{04}(T^7 + T^5 - 23T^3 + 9T) \\
 & - \frac{1}{1296} \gamma_{03}^3(T^9 + 6T^7 + 6T^5 - 150T^3 + 41T) - \frac{1}{720} \gamma_{06}(T^6 - 10T^4 + 15T^2) \\
 & + \frac{1}{720} \gamma_{03} \gamma_{05}(T^8 - 52T^4 + 72T^2 + 3) + \frac{1}{1152} \gamma_{04}^2(T^8 + 2T^6 - 57T^4 + 78T^2) \\
 & - \frac{1}{1728} \gamma_{03}^2 \gamma_{04}(T^{10} + 8T^8 - 4T^6 - 434T^4 + 513T^2 + 24) \\
 & \left. + \frac{1}{31104} \gamma_{03}^4(T^{12} + 14T^{10} + 93T^8 - 96T^6 - 4501T^4 + 4514T^2 + 275) \right], \quad \dots \quad (6.9)
 \end{aligned}$$

the earlier terms of which appeared as equation (24) in the 1956 paper.

For a specified cut-off proportion, P , equation (5.1) may be written

$$M_X(\theta) = P^{-1} \int_U^\infty F(D) Z(u) du. \quad \dots \quad (6.10)$$

Since

$$\int_U^\infty D^r Z(u) du = -D^{r-1} Z(u) \text{ at } u = U, \quad \dots \quad (6.11)$$

terms on the right hand side of (6.10) can be expressed in Hermite polynomials with argument U . If the selection is specified by η , the point of cut-off, the moments of X are now available. These were not presented in the previous paper, but selection by a pre-determined cut-off level may often be employed and the first four moments are, therefore, given below in equations (6.13) to $0(n^{-2})$ and (6.15), (6.17), (6.19) to $0(n^{-3/2})$. Each of these involves $P(U)$, which can be determined from the terms independent of θ in (6.10) in the form of $P_X(U)$ and an adjustment consisting of an infinite series of Hermite polynomials multiplied by $Z(U)$; this appears as equation (6.12). An alternative form of results is obtained by substitution from (6.8) and (6.9), so expressing the moments with argument T instead of U . Each moment is then written as a series involving the cumulants of $p(x, y)$, or more conveniently the γ_{ij} , and T , the unit normal deviate corresponding to the proportion selected, P . These results, generalizations of equations (26), (28), (29), (30) in the previous paper, appear as (6.14), (6.16), (6.18), and (6.20) below. In all these formulae, the grouping of terms retains the pattern of the Cornish-Fisher expansion.

THE TRANSFORMATION OF A DISTRIBUTION UNDER SELECTION

From the terms independent of θ in (6.10),

$$P(U) = P_N(U) + \left(\frac{1}{6} \gamma_{03} D^2 - \frac{1}{24} \gamma_{04} D^3 - \frac{1}{72} \gamma_{03}^2 D^5 + \frac{1}{120} \gamma_{05} D^4 \right. \\ \left. + \frac{1}{144} \gamma_{03} \gamma_{04} D^6 + \frac{1}{1296} \gamma_{03}^3 D^8 \right) Z(u) \quad \text{at } u = U. \quad \dots (6.12)$$

The terms in $(\sigma\theta)$ have been evaluated to order n^{-2} , as accurate values for the mean are likely to be particularly important. The results are

$$\mu'_1(X) = P^{-1} \sigma \left[\rho - \frac{1}{6} (3\gamma_{12} D + \rho \gamma_{03} D^3) + \frac{1}{24} (4\gamma_{13} D^2 + \rho \gamma_{04} D^4) + \frac{1}{72} (6\gamma_{03} \gamma_{12} D^4 + \rho \gamma_{03}^2 D^6) \right. \\ - \frac{1}{120} (5\gamma_{14} D^3 + \rho \gamma_{05} D^5) - \frac{1}{144} (4\gamma_{03} \gamma_{13} D^5 + 3\gamma_{12} \gamma_{04} D^5 + \rho \gamma_{03} \gamma_{04} D^7) \\ - \frac{1}{1296} (9\gamma_{03}^2 \gamma_{12} D^7 + \rho \gamma_{03}^3 D^9) + \frac{1}{720} (6\gamma_{15} D^4 + \rho \gamma_{06} D^6) \\ + \frac{1}{720} (5\gamma_{03} \gamma_{14} D^6 + 3\gamma_{12} \gamma_{05} D^6 + \rho \gamma_{03} \gamma_{05} D^8) + \frac{1}{1152} (8\gamma_{04} \gamma_{13} D^6 + \rho \gamma_{04}^2 D^8) \\ + \frac{1}{1728} (6\gamma_{03} \gamma_{12} \gamma_{04} D^8 + 4\gamma_{03}^2 \gamma_{13} D^8 + \rho \gamma_{03}^2 \gamma_{04} D^{10}) \\ \left. + \frac{1}{31104} (12\gamma_{03}^3 \gamma_{12} D^{10} + \rho \gamma_{03}^4 D^{12}) \right] Z(u) \quad \text{at } u = U \quad \dots (6.13)$$

and

$$\mu'_1(X) = \frac{\sigma Z(T)}{P} \left[\rho + \frac{1}{6} (3\gamma_{12} - 2\rho \gamma_{03}) T + \frac{1}{24} (4\gamma_{13} - 3\rho \gamma_{04}) (T^2 - 1) \right. \\ - \frac{1}{36} (6\gamma_{03} \gamma_{12} - 5\rho \gamma_{03}^2) (2T^2 - 1) \\ + \frac{1}{120} (5\gamma_{14} - 4\rho \gamma_{05}) (T^3 - 3T) - \frac{1}{24} (4\gamma_{03} \gamma_{13} + 3\gamma_{12} \gamma_{04} - 6\rho \gamma_{03} \gamma_{04}) (T^3 - 2T) \\ + \frac{1}{324} (9\gamma_{03}^2 \gamma_{12} - 8\rho \gamma_{03}^2) (12T^3 - 17T) + \frac{1}{720} (6\gamma_{15} - 5\rho \gamma_{06}) (T^4 - 6T^2 + 3) \\ - \frac{1}{180} (5\gamma_{03} \gamma_{14} + 3\gamma_{12} \gamma_{05} - 7\rho \gamma_{03} \gamma_{05}) (2T^4 - 9T^2 + 3) \\ - \frac{1}{384} (8\gamma_{04} \gamma_{13} - 7\rho \gamma_{04}^2) (3T^4 - 12T^2 + 5) \\ + \frac{1}{288} (6\gamma_{03} \gamma_{12} \gamma_{04} + 4\gamma_{03}^2 \gamma_{13} - 9\rho \gamma_{03}^2 \gamma_{04}) (14T^4 - 47T^2 + 13) \\ \left. - \frac{1}{7776} (12\gamma_{03}^3 \gamma_{12} - 11\rho \gamma_{03}^4) (252T^4 - 680T^2 + 151) \right]. \quad \dots (6.14)$$

Similarly, from the coefficient of $(\sigma\theta)^2$,

$$\begin{aligned}\mu'_2(X^2) = & \sigma^2 + P^{-1}\sigma^2 \left[\rho^2 D - \frac{1}{6} \{6\gamma_{21} + (\gamma_{03} + 6\rho\gamma_{12})D^2 + \rho^2\gamma_{03}D^4\} \right. \\ & + \frac{1}{24} \{12\gamma_{22}D + (\gamma_{04} + 8\rho\gamma_{13})D^3 + \rho^2\gamma_{04}D^5\} \\ & + \frac{1}{72} \{(12\gamma_{03}\gamma_{21} + 18\gamma_{12}^2)D^3 + (\gamma_{03}^2 + 12\rho\gamma_{03}\gamma_{12})D^5 + \rho^2\gamma_{03}^2D^7\} \\ & - \frac{1}{120} \{20\gamma_{23}D^2 + (\gamma_{05} + 10\rho\gamma_{14})D^4 + \rho^2\gamma_{05}D^6\} \\ & - \frac{1}{144} \{(12\gamma_{03}\gamma_{22} + 24\gamma_{12}\gamma_{13} + 6\gamma_{21}\gamma_{04})D^4 \\ & \quad + (\gamma_{03}\gamma_{04} + 8\rho\gamma_{03}\gamma_{13} + 6\rho\gamma_{12}\gamma_{04})D^6 + \rho^2\gamma_{03}\gamma_{04}D^8\} \\ & - \frac{1}{1296} \{(18\gamma_{03}^2\gamma_{21} + 54\gamma_{03}\gamma_{12}^2)D^6 + (\gamma_{03}^3 + 18\rho\gamma_{03}^2\gamma_{12})D^8 \\ & \quad \left. + \rho^2\gamma_{03}^3D^{10}\} \right] Z(u) \quad \text{at } u = U \quad \dots (6.15)\end{aligned}$$

and, with the aid of (6.12),

$$\begin{aligned}\mu'_2(X^2) = & \sigma^2 + \frac{\sigma^2 Z(T)}{P} \left[\rho^2 T + \frac{1}{3} \{3\gamma_{21} + 3\rho\gamma_{12}(T^2 - 1) - \rho^2\gamma_{03}(2T^2 - 1)\} \right. \\ & + \frac{1}{12} \{6\gamma_{22}T + 4\rho\gamma_{13}(T^3 - 3T) - 3\rho^2\gamma_{04}(T^3 - 2T)\} \\ & + \frac{1}{36} \{-12\gamma_{03}\gamma_{21}T + 9\gamma_{12}^2(T^3 - 3T) - 36\rho\gamma_{03}\gamma_{12}(T^3 - 2T) + 2\rho^2\gamma_{03}^2(12T^3 - 17T)\} \\ & + \frac{1}{60} \{10\gamma_{23}(T^2 - 1) + 5\rho\gamma_{14}(T^4 - 6T^2 + 3) - 2\rho^2\gamma_{05}(2T^4 - 9T^2 + 3)\} \\ & + \frac{1}{72} \{-12\gamma_{03}\gamma_{22}(2T^2 - 1) + 12\gamma_{12}\gamma_{13}(T^4 - 6T^2 + 3) - 9\gamma_{21}\gamma_{04}(T^2 - 1) \\ & \quad - 16\rho\gamma_{03}\gamma_{13}(2T^4 - 9T^2 + 3) - 9\rho\gamma_{12}\gamma_{04}(3T^4 - 12T^2 + 5) \\ & \quad + 3\rho^2\gamma_{03}\gamma_{04}(14T^4 - 47T^2 + 13)\} \\ & + \frac{1}{324} \{45\gamma_{03}^2\gamma_{21}(2T^2 - 1) - 54\gamma_{03}\gamma_{12}^2(2T^4 - 9T^2 + 3) \\ & \quad \left. + 27\rho\gamma_{03}^2\gamma_{12}(14T^4 - 47T^2 + 13) - \rho^2\gamma_{03}^3(252T^4 - 680T^2 + 151)\} \right]. \quad (6.16)\end{aligned}$$

THE TRANSFORMATION OF A DISTRIBUTION UNDER SELECTION

Again, from the coefficient of $(\sigma\theta)^3$,

$$\begin{aligned} \mu'_3(X^3) = & \sigma^3\gamma_{30} + P^{-1}\sigma^3 \left[3\rho + \rho^3 D^2 - \frac{1}{6} \{ (9\gamma_{12} + 18\rho\gamma_{21})D + (3\rho\gamma_{03} + 9\rho^2\gamma_{12})D^3 + \rho^3\gamma_{03}D^5 \} \right. \\ & + \frac{1}{24} \{ 24\gamma_{31} + (12\gamma_{13} + 36\rho\gamma_{22})D^2 + (3\rho\gamma_{04} + 12\rho^2\gamma_{13})D^4 + \rho^3\gamma_{04}D^6 \} \\ & + \frac{1}{72} \{ 108\gamma_{12}\gamma_{21}D^2 + (18\gamma_{03}\gamma_{12} + 36\rho\gamma_{03}\gamma_{21} + 54\rho\gamma_{12}^2)D^4 \\ & \quad + (3\rho\gamma_{03}^2 + 18\rho^2\gamma_{03}\gamma_{12})D^6 + \rho^3\gamma_{03}^2D^8 \} \\ & - \frac{1}{120} \{ 60\gamma_{32}D + (15\gamma_{14} + 60\rho\gamma_{23})D^3 + (3\rho\gamma_{05} + 15\rho^2\gamma_{14})D^5 + \rho^3\gamma_{05}D^7 \} \\ & - \frac{1}{144} \{ (24\gamma_{03}\gamma_{31} + 108\gamma_{12}\gamma_{22} + 72\gamma_{21}\gamma_{13})D^3 \\ & \quad + (12\gamma_{03}\gamma_{13} + 9\gamma_{12}\gamma_{04} + 36\rho\gamma_{03}\gamma_{22} + 72\rho\gamma_{12}\gamma_{13} + 18\rho\gamma_{21}\gamma_{04})D^5 \\ & \quad + (3\rho\gamma_{03}\gamma_{04} + 12\rho^2\gamma_{03}\gamma_{13} + 9\rho^2\gamma_{12}\gamma_{04})D^7 + \rho^3\gamma_{03}\gamma_{04}D^9 \} \\ & - \frac{1}{1296} \{ (324\gamma_{03}\gamma_{12}\gamma_{21} + 162\gamma_{12}^3)D^5 \\ & \quad + (27\gamma_{03}^2\gamma_{12} + 54\rho\gamma_{03}^2\gamma_{21} + 162\rho\gamma_{03}\gamma_{12}^2)D^7 \\ & \quad + (3\rho\gamma_{03}^3 + 27\rho^2\gamma_{03}^2\gamma_{12})D^9 \\ & \quad \left. + \rho^3\gamma_{03}^3D^{11} \right] Z(u) \text{ at } u = U \end{aligned} \quad \dots \quad (6.17)$$

and

$$\begin{aligned} \mu'_3(X^3) = & \sigma^3\gamma_{30} + \frac{\sigma^3 Z(T)}{P} \left[3\rho + \rho^3(T^2 - 1) + \frac{1}{2} \{ 3(\gamma_{12} + 2\rho\gamma_{21})T - 2\rho\gamma_{03}T \right. \\ & \quad + 3\rho^2\gamma_{12}(T^3 - 3T) - 2\rho^3\gamma_{03}(T^3 - 2T) \} \\ & + \frac{1}{8} \{ 8\gamma_{31} + 4(\gamma_{13} + 3\rho\gamma_{22})(T^2 - 1) - 3\rho\gamma_{04}(T^2 - 1) + 4\rho^2\gamma_{13}(T^4 - 6T^2 + 3) \\ & \quad - \rho^3\gamma_{04}(3T^4 - 12T^2 + 5) \} \\ & + \frac{1}{12} \{ 18\gamma_{12}\gamma_{21}(T^2 - 1) - 6(\gamma_{03}\gamma_{12} + 2\rho\gamma_{03}\gamma_{21})(2T^2 - 1) + 9\rho\gamma_{12}^2(T^4 - 6T^2 + 3) \\ & \quad + 5\rho\gamma_{03}^2(2T^2 - 1) - 12\rho^2\gamma_{03}\gamma_{12}(2T^4 - 9T^2 + 3) + \rho^3\gamma_{03}^2(14T^4 - 47T^2 + 13) \} \\ & + \frac{1}{40} \{ 20\gamma_{32}T + 5(\gamma_{14} + 4\rho\gamma_{23})(T^3 - 3T) - 4\rho\gamma_{05}(T^3 - 3T) \\ & \quad + 5\rho^2\gamma_{14}(T^5 - 10T^3 + 15T) - 4\rho^3\gamma_{05}(T^5 - 7T^3 + 8T) \} \\ & + \frac{1}{24} \{ -8\gamma_{03}\gamma_{31}T + 6(3\gamma_{12}\gamma_{22} + 2\gamma_{21}\gamma_{13})(T^3 - 3T) \\ & \quad - 3(4\gamma_{03}\gamma_{13} + 3\gamma_{12}\gamma_{04} + 12\rho\gamma_{03}\gamma_{22} + 6\rho\gamma_{31}\gamma_{04})(T^3 - 2T) \\ & \quad + 12\rho\gamma_{12}\gamma_{13}(T^5 - 10T^3 + 15T) + 18\rho\gamma_{03}\gamma_{04}(T^3 - 2T) \\ & \quad - 20\rho^2\gamma_{03}\gamma_{13}(T^5 - 8T^3 + 9T) - 18\rho^2\gamma_{12}\gamma_{04}(T^5 - 7T^3 + 8T) \\ & \quad + 2\rho^3\gamma_{03}\gamma_{04}(12T^5 - 69T^3 + 64T) \} \\ & + \frac{1}{216} \{ -324\gamma_{03}\gamma_{12}\gamma_{21}(T^3 - 2T) + 27\gamma_{12}^3(T^5 - 10T^3 + 15T) \\ & \quad + 18(\gamma_{03}^2\gamma_{12} + 2\rho\gamma_{03}^2\gamma_{21})(12T^3 - 17T) - 270\rho\gamma_{03}\gamma_{12}^2(T^5 - 8T^3 + 9T) \\ & \quad - 16\rho\gamma_{03}^3(12T^3 - 17T) + 36\rho^2\gamma_{03}^2\gamma_{12}(16T^5 - 101T^3 + 90T) \\ & \quad - 4\rho^3\gamma_{03}^3(80T^5 - 403T^3 + 304T) \} \left. \right] \end{aligned} \quad \dots \quad (6.18)$$

Finally, from $(\sigma\theta)^4$,

$$\mu'_4(X^4) = \sigma^4(3 + \gamma_{40})$$

$$\begin{aligned} & -P^{-1}\sigma^4 \left[6\rho^2 D + \rho^4 D^3 - \frac{1}{6} \{ (36\gamma_{21} + 24\rho\gamma_{30} + (36\rho\gamma_{12} + 36\rho^2\gamma_{21})D^2 \right. \\ & \qquad \qquad \qquad \left. + (6\rho^2\gamma_{03} + 12\rho^3\gamma_{12})D^4 + \rho^4\gamma_{03}D^6 \} \right. \\ & + \frac{1}{24} \{ (72\gamma_{22} + 96\rho\gamma_{31})D + (48\rho\gamma_{13} + 72\rho^2\gamma_{22})D^3 + (6\rho^2\gamma_{04} + 16\rho^3\gamma_{13})D^5 + \rho^4\gamma_{04}D^7 \} \\ & + \frac{1}{72} \{ (144\gamma_{12}\gamma_{30} + 216\gamma_{21}^2)D + (72\gamma_{03}\gamma_{21} + 108\gamma_{12}^2 + 48\rho\gamma_{03}\gamma_{30} + 432\rho\gamma_{12}\gamma_{21})D^3 \\ & \qquad \qquad \qquad + (72\rho\gamma_{03}\gamma_{12} + 72\rho^2\gamma_{03}\gamma_{21} + 108\rho^2\gamma_{12}^2)D^5 \\ & \qquad \qquad \qquad + (6\rho^2\gamma_{03}^2 + 24\rho^3\gamma_{03}\gamma_{12})D^7 + \rho^4\gamma_{03}^2D^9 \} \\ & - \frac{1}{120} \{ 120\gamma_{41} + (120\gamma_{23} + 240\rho\gamma_{32})D^2 + (60\rho\gamma_{14} + 120\rho^2\gamma_{23})D^4 \\ & \qquad \qquad \qquad + (6\rho^2\gamma_{05} + 20\rho^3\gamma_{14})D^6 + \rho^4\gamma_{05}D^8 \} \\ & - \frac{1}{144} \{ (288\gamma_{12}\gamma_{31} + 432\gamma_{21}\gamma_{22} + 96\gamma_{30}\gamma_{13})D^2 \\ & \qquad \qquad \qquad + (72\gamma_{03}\gamma_{22} + 144\gamma_{12}\gamma_{13} + 36\gamma_{21}\gamma_{04} + 96\rho\gamma_{03}\gamma_{31} + 432\rho\gamma_{12}\gamma_{22} \\ & \qquad \qquad \qquad + 288\rho\gamma_{21}\gamma_{13} + 24\rho\gamma_{30}\gamma_{04})D^4 \\ & \qquad \qquad \qquad + (48\rho\gamma_{03}\gamma_{13} + 72\rho^2\gamma_{03}\gamma_{22} + 36\rho\gamma_{12}\gamma_{04} + 144\rho^2\gamma_{12}\gamma_{13} + 36\rho^2\gamma_{21}\gamma_{04})D^6 \\ & \qquad \qquad \qquad + (6\rho^2\gamma_{03}\gamma_{04} + 16\rho^3\gamma_{03}\gamma_{13} + 12\rho^3\gamma_{12}\gamma_{04})D^8 + \rho^4\gamma_{03}\gamma_{04}D^{10} \} \\ & - \frac{1}{1296} \{ (432\gamma_{03}\gamma_{12}\gamma_{30} + 648\gamma_{03}\gamma_{21}^2 + 1944\gamma_{12}^2\gamma_{21})D^4 \\ & \qquad \qquad \qquad + (108\gamma_{03}^2\gamma_{21} + 324\gamma_{03}\gamma_{12}^2 + 72\rho\gamma_{03}^2\gamma_{30} + 1296\rho\gamma_{03}\gamma_{12}\gamma_{21} + 648\rho\gamma_{12}^3)D^6 \\ & \qquad \qquad \qquad + (108\rho\gamma_{03}^2\gamma_{12} + 108\rho^2\gamma_{03}^2\gamma_{21} + 324\rho^2\gamma_{03}\gamma_{12}^2)D^8 \\ & \qquad \qquad \qquad + (6\rho^2\gamma_{03}^3 + 36\rho^3\gamma_{03}^2\gamma_{12})D^{10} \\ & \qquad \qquad \qquad + \rho^4\gamma_{03}^3D^{12} \} \Big] Z(u) \text{ at } u = U \end{aligned}$$

... (6.19)

and

$$\mu'_4(X^4) = \sigma^4(3 + \gamma_{40})$$

$$\begin{aligned} & + \frac{\sigma^4 Z(T)}{P} \left[6\rho^2 T + \rho^4 (T^3 - 3T) \right. \\ & + \frac{1}{3} \{ 6(3\gamma_{21} + 2\rho\gamma_{30}) + 18(\rho\gamma_{12} + \rho^2\gamma_{21})(T^2 - 1) \\ & \qquad \qquad \qquad - 6\rho^2\gamma_{03}(2T^2 - 1) + 6\rho^3\gamma_{12}(T^4 - 6T^2 + 3) \\ & \qquad \qquad \qquad - 2\rho^4\gamma_{03}(2T^4 - 9T^2 + 3) \} \\ & + \frac{1}{6} \{ 6(3\gamma_{22} + 4\rho\gamma_{31})T + 6(2\rho\gamma_{13} + 3\rho^2\gamma_{22})(T^3 - 3T) \\ & \qquad \qquad \qquad - 9\rho^2\gamma_{04}(T^3 - 2T) + 4\rho^3\gamma_{13}(T^5 - 10T^3 + 15T) \\ & \qquad \qquad \qquad - 3\rho^4\gamma_{04}(T^5 - 7T^3 + 8T) \} \\ & + \frac{1}{18} \{ 18(2\gamma_{12}\gamma_{30} + 3\gamma_{21}^2)T - 12(3\gamma_{03}\gamma_{21} + 2\rho\gamma_{03}\gamma_{30})T + 27(\gamma_{12}^2 + 4\rho\gamma_{12}\gamma_{21})(T^3 - 3T) \\ & \qquad \qquad \qquad - 108(\rho\gamma_{03}\gamma_{12} + \rho^2\gamma_{03}\gamma_{21})(T^3 - 2T) + 27\rho^2\gamma_{12}^2(T^5 - 10T^3 + 15T) \\ & \qquad \qquad \qquad + 4\rho^2\gamma_{03}^2(12T^3 - 17T) - 60\rho^3\gamma_{03}\gamma_{12}(T^5 - 8T^3 + 9T) \\ & \qquad \qquad \qquad + 2\rho^4\gamma_{03}^2(16T^5 - 101T^3 + 90T) \} \end{aligned}$$

THE TRANSFORMATION OF A DISTRIBUTION UNDER SELECTION

$$\begin{aligned}
 & + \frac{1}{30} \{ 30\gamma_{41} + 30(\gamma_{23} + 2\rho\gamma_{32})(T^2 - 1) + 15(\rho\gamma_{14} + 2\rho^2\gamma_{23})(T^4 - 6T^2 + 3) \\
 & \quad - 6\rho^2\gamma_{05}(2T^4 - 9T^2 + 3) + 5\rho^3\gamma_{14}(T^6 - 15T^4 + 45T^2 - 15) \\
 & \quad - 2\rho^4\gamma_{05}(2T^6 - 21T^4 + 48T^2 - 12) \} \\
 & + \frac{1}{36} \{ 12(6\gamma_{12}\gamma_{31} + 9\gamma_{21}\gamma_{22} + 2\gamma_{30}\gamma_{13})(T^2 - 1) \\
 & \quad - 12(3\gamma_{03}\gamma_{22} + 4\rho_{03}\gamma_{31})(2T^2 - 1) - 9(3\gamma_{21}\gamma_{04} + 2\rho\gamma_{30}\gamma_{04})(T^2 - 1) \\
 & \quad + 36(\gamma_{12}\gamma_{13} + 3\rho\gamma_{12}\gamma_{22} + 2\gamma_{21}\gamma_{13})(T^4 - 6T^2 + 3) \\
 & \quad - 24(2\rho\gamma_{03}\gamma_{13} + 3\rho^2\gamma_{03}\gamma_{22})(2T^4 - 9T^2 + 3) \\
 & \quad - 27(\rho\gamma_{12}\gamma_{04} + \rho^2\gamma_{04}\gamma_{21})(3T^4 - 12T^2 + 5) \\
 & \quad + 36\rho^2\gamma_{12}\gamma_{13}(T^6 - 15T^4 + 45T^2 - 15) \\
 & \quad + 9\rho^2\gamma_{03}\gamma_{04}(14T^4 - 47T^2 + 13) - 24\rho^3\gamma_{03}\gamma_{13}(2T^6 - 25T^4 + 60T^2 - 15) \\
 & \quad - 45\rho^3\gamma_{12}\gamma_{04}(T^6 - 11T^4 - 25T^2 - 7) \\
 & \quad + 3\rho^4\gamma_{03}\gamma_{04}(18T^6 - 165T^4 + 312T^2 - 67) \} \\
 & + \frac{1}{162} \{ -54(2\gamma_{03}\gamma_{12}\gamma_{30} + 3\gamma_{03}\gamma_{21}^2)(2T^2 - 1) + 243\gamma_{12}^2\gamma_{21}(T^4 - 6T^2 + 3) \\
 & \quad + 45(3\gamma_{03}^2\gamma_{21} + 2\rho\gamma_{03}^2\gamma_{30})(2T^2 - 1) - 162(\gamma_{03}\gamma_{12}^2 + 4\rho\gamma_{03}\gamma_{12}\gamma_{21})(2T^4 - 9T^2 + 3) \\
 & \quad + 81\rho\gamma_{12}^3(T^6 - 15T^4 + 45T^2 - 15) \\
 & \quad + 81(\rho\gamma_{03}^2\gamma_{12} + \rho^2\gamma_{03}^2\gamma_{21})(14T^4 - 47T^2 + 13) \\
 & \quad - 243\rho^2\gamma_{03}\gamma_{12}^2(2T^6 - 25T^4 + 60T^2 - 15) \\
 & \quad - 3\rho^2\gamma_{03}^3(252T^4 - 680T^2 + 151) \\
 & \quad + 45\rho^3\gamma_{03}^2\gamma_{12}(18T^6 - 185T^4 + 360T^2 - 75) \\
 & \quad - \rho^4\gamma_{03}^3(396T^6 - 3332T^4 + 5427T^2 - 957) \} \dots \quad (6.20)
 \end{aligned}$$

Moments about zero alone have been given, as apparently no appreciable reductions of formulae occur when the higher moments are expressed in deviations from $\mu'_1(X)$. Equations (6.14), (6.16), (6.18) and (6.20) reduce to (26), (28), (29), and (30) of Finney (1956) (except that additional terms are now available), by substitution of the parameters particularly appropriate to the earlier model:

$$\left. \begin{aligned}
 \rho &= \frac{\sigma}{\omega}, \\
 \frac{\sigma^3}{\omega^3} \gamma_{30} &= \frac{\sigma^2}{\omega^2} \gamma_{21} = \frac{\sigma}{\omega} \gamma_{12} = \gamma_{03} = c, \\
 \frac{\sigma^4}{\omega^4} \gamma_{40} &= \frac{\sigma^3}{\omega^3} \gamma_{31} = \frac{\sigma^2}{\omega^2} \gamma_{22} = \frac{\sigma}{\omega} \gamma_{13} = \gamma_{04} = d, \\
 \frac{\sigma^5}{\omega^5} \gamma_{50} &= \frac{\sigma^4}{\omega^4} \gamma_{41} = \frac{\sigma^3}{\omega^3} \gamma_{32} = \frac{\sigma^2}{\omega^2} \gamma_{23} = \frac{\sigma}{\omega} \gamma_{14} = \gamma_{05} = e.
 \end{aligned} \right\} \dots \quad (6.21)$$

7. TABLES

In the use of the results of this paper, numerous functions of the unit normal deviate, T , are required. Possibly the most commonly needed are those involved in equations (4.16) and (6.14). The following functions :

$$T, Z(T), \frac{d^i}{dT^i} \log P_N(T) \quad \text{for } i = 1, 2, 3, 4, 5,$$

$T/6, (T^2-1)/24, -(2T^2-1)/36, (T^3-3T)/120, -(T^3-2T)/24, (12T^3-17T)/324$ were tabulated in Finney (1956) for $P_N(T) = 0.001, 0.005, 0.01, 0.0125, 0.02, 0.025 (0.025) 0.100(0.05) 0.95$. This selection of values was chosen as containing the most useful for preliminary studies of 2-stage selection.

Since then, Mr. S. Lipton (University of New South Wales) has prepared extended tables of the same functions, so far for the range $P_N(T) = 0.00001 (0.00001) 0.00250, 0.0001 (0.0001) 0.0250, 0.001(0.001) 0.500$. These have not yet been published.

I am indebted to Mr. A. D. Henderson for checking part of the algebra.

REFERENCES

- BARTLETT, M. S. (1938a): The characteristic function of a conditional statistic. *Journal of the London Mathematical Society*, **13**, 62-67.
- (1938b): Further aspects of the theory of multiple regression. *Proceedings of the Cambridge Philosophical Society*, **34**, 33-40.
- COCHRAN, W. G. (1951): Improvement by means of selection. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 449-470.
- CORNISH, E. A. and FISHER, R. A. (1937): Moments and cumulants in the specification of distributions. *Revue de l'Institut International de Statistique*, **5**, 307-320.
- CURNOW, R. N. (1960): The regression of true value on estimated value. *Biometrika* (in Press).
- FINNEY, D. J. (1956): The consequences of selection for a variate subject to errors of measurement. *Revue de l'Institut International de Statistique*, **24**, 1-10.
- (1958): Statistical problems of plant selection. *Bulletin de l'Institut International de Statistique*, **36**(3), 242-268.
- FISHER, R. A. and CORNISH, E. A. (1960): The percentile points of distributions having known cumulants. *Technometrics*, **2**, 209-225.
- LAWLEY, D. N. (1943): A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh*, **62**, 28-30.
- LINDLEY, D. V. (1947): Regression lines and the linear functional relationship. *J. Roy. Stat. Soc., Supplement*, **9**, 218-249.
- PEARSON, K. (1912): On the general theory of the influence of selection on correlation and variation. *Biometrika*, **8**, 437-443.
- ROBERTSON, A. (1957): Optimum group size in progeny testing and family selection. *Biometrics*, **13**, 442-450.

Paper received: March, 1961.

ANALYSIS OF ERRORS IN CENSUSES AND SURVEYS WITH SPECIAL REFERENCE TO EXPERIENCE IN INDIA*

By P. C. MAHALANOBIS and D. B. LAHIRI

Indian Statistical Institute

I. INTRODUCTION

1. The demand for statistical information is growing increasingly and rapidly, and survey organisations are hard pressed to supply results with the required speed. They must assess the quality of the data and caution the users that the results are not perfect, but they must avoid creating an undue impression that the data are so unreliable as to be of no use. The manner in which the existence of errors can be presented would depend on the level of statistical maturity of the country, not only of a select group of survey experts but also of the direct users, the policy makers, and to some extent the general public who may be affected by any policy decision based (at least partially) on the statistical information. There, however, is general agreement among survey experts that survey reports should supply some idea of the reliability of the results.

2. The study of errors in surveys has two purposes. First, to guide the user in interpreting the results. Secondly, in improving the quality of future surveys. Even when great care has been taken to set up important controls it is still necessary to get an assurance that the controls were effective and results with desired accuracy¹ have been obtained.

3. In India a number of techniques are being used for evaluation of survey results. All depend on comparisons among alternative (independent) estimates. Some of these may on *a priori* grounds be assumed to be more reliable than the others.

4. An illustration of this is provided in the evaluation of the results supplied by (a) (so-called) complete enumeration and (b) by sample survey of the outturn of jute in Bengal against (c) statistics of jute trade obtained subsequently which are known to be very reliable. This is a rather rare example where opportunity had occurred of comparing three sets of estimates for two consecutive years.

5. In some situations there are no *a priori* grounds for preferring one estimate to another. In this case the divergence between two or more independent estimates may provide a basis for a mental appraisal of the margin of uncertainty. This is

* This paper originally presented at the 32nd Session of the International Statistical Institute is being reprinted here with the kind permission of the Editor, *Bulletin of the International Statistical Institute*.

¹ For practical purposes the desired accuracy can be often thought of or specified as the margin of 'permissible error' in the sense that any decision to be taken on the basis of sample estimates would remain the same within the limits of the permissible error.

made possible if the results are based on interpenetrating net-work of (sub) samples (IPNS) each of which is surveyed and/or processed by different but comparable operational units. When two (or more) samples are drawn from the same population and covered according to the same survey design, the results based on the different samples are equally valid, even though they are derived by different operational units; and divergences between the different sets of estimates supply directly some idea of the margin of uncertainty. It will be noticed that the operational differentials find a reflection in this manner of assessing the margin of uncertainty.

6. The interpenetration at the operational level may be of various types. For example, the field work is conducted by the same agency but the processing is done by different agencies.² Or, the field work is by different agencies but processing by the same agency; for example, in the survey of landholdings in India (reported in this paper) the entire processing work was done in the Indian Statistical Institute, even though the field work was conducted by both the Central and State agencies. Again, there may be complete interpenetration where both the field and processing work are done by completely different agencies; this is the form which the Indian National Sample Survey (NSS) is gradually assuming with the participation of the different States of India. Even for the same agency our normal practice is to arrange field work with interpenetration in respect of parties of investigators. Study of party and agency differences finds a place in our section on landholdings. There are also some marginal cases where even for the same agency the processing is done at different centres, or by different tabulating units, or at different points of time.

7. It should, however, be pointed out that, strictly speaking, a complete absence of *a priori* preferences is not always a reality. For example, if it is known that a particular agency has had a long experience in a particular field then the results thrown up by it may be accepted to have higher validity. Or, when one survey is carried out by temporary *ad hoc* staff and another survey by a whole-time permanent statistical staff then some may be inclined to accept the results thrown up by the latter to have higher validity. For the same reason one may accept an intensively supervised well conducted sample survey by qualified, experienced and well-trained investigators to have higher validity compared to a complete enumeration conducted under usual census conditions. Examples of this situation will be found in our sections on spot-check of crop census records and sample verification of livestock census.

8. One difficulty of evaluation is that the assessor may not be in possession of full background information about the agency, or about the conditions under which the census data, for example, may have been collected or processed. Sometimes comparisons between census and sample check or those between interpenetrating

² This is the practice for the pre-harvest crop-acreage survey where the field work is conducted by *patwaris* under the State authorities. The processing is done by three agencies each covering two sub-samples, the agencies being (a) the State authorities, (b) the Directorate of National Sample Survey, Government of India, and (c) the Indian Statistical Institute.

EVALUATION OF STATISTICAL SURVEYS

samples covered by different agencies may provide corroborative evidence in support of certain 'feelings' based on previously available background information which may even be of a vague and inadequate nature.

9. It may be worth-while stressing an obvious point which, however, is sometimes not kept in mind. For proper evaluation of results it is essential to take into consideration all available information, not merely the internal evidences supplied by the survey alone; not merely the quantitative external evidences but also even vague non-quantitative evidences. There is usually a residual element of subjective appreciation even when decisions are sought to be made in an objective way. The aim must be to utilise the whole of the available evidence in such a way that a maximum amount of agreement can be reached among competent assessors. This is the basic approach of science.

10. In some methods of evaluation the emphasis is on unitary check where the purpose is to evaluate the quality of data collected at the primary unit of enquiry. Evaluation of results for small administrative units is also possible. These are illustrated by the spot-check of crop census records and sample verification of the livestock census. Such checks can supply not only material for the study of errors of ascertainment, unit by unit, but if performed on a random basis, may also provide a means for separation of total *net* error of an aggregate estimate into ascertainment error and errors of coverage and compilation. (See our section on livestock census).

11. Another useful technique is to break up the survey period, for example, one round³ of the National Sample Survey (NSS) into a number of sub-rounds, and to compare the estimates for each sub-round. Differences in such sub-round estimates, if any, would give valuable information for proper interpretation of the data. This is an example of a survey plan where the work of the same agency can be evaluated against its own work conducted under somewhat different conditions. (See our section on population growth).

12. The above idea of self-evaluation has been used in other forms. For example, in crop cutting experiments, to estimate the yield of crops per hectare, the crop is harvested separately, at each sample point, in the form of two or three concentric sample cuts. To what extent the work has been done under control can be then studied from the magnitude of the divergence between different estimates, each based on a different size of cut.

13. A similar device which has been found to be of value, is using more than one reference period of time for the collection of data by the interview method. In some designs the data are collected for a long period (for example, one year) in such a way that tabulations by shorter reference periods (e.g. one month) are possible. Comparisons of results based on different periods of reference may reveal factors

³ NSS is a multi-purpose survey, a specified group of subjects is covered in an integrated manner in a single "round" to be completed in a specified survey-period; different rounds have usually a different survey plan depending upon the group of subjects chosen and upon the relative emphasis placed on different aspects of the survey.

like 'recall lapse.' An illustration of this is provided in the survey design for the estimation of birth, death and growth rates.

14. Another technique is to compare estimates obtained from one method of collecting the data against another more reliable method for estimating not the same character but a related one which sets a lower (or upper) limit to the former character. An illustration is provided in the comparison of death rates obtained by the interview method against a method of keeping a watch (by the method of re-enumeration) on a sample of individuals. Comparisons may also be made of results obtained by a less intensive enquiry against those based on a more intensive one. (See section on population growth).

15. As the Indian National Sample Survey (NSS) is carrying out surveys round after round, it is possible to compare the results based on two or more rounds.⁴ 'Consistency' over rounds adds to the confidence with which the results may be accepted.

16. Another technique which appears to have good potentialities is comparison over 'space' as distinguished from 'time' (or rounds) which we have just described. Illustrations of similarity of the nature of divergence from one area to another are to be found in our sections on landholdings and sample verification of livestock census.

II. JUTE PRODUCTION IN BENGAL 1944-45 AND 1945-46

17. It is not always that one gets an opportunity of evaluating the results obtained on the basis of complete enumeration and sample survey against a third, but extremely reliable, figure. Such an opportunity was availed of in two consecutive years in regard to the 1944-45 and 1945-46 Jute Crop of Bengal.

18. Jute being a cash crop of international importance, accurate export trade figures are maintained and become available about 15 months after the harvesting season. Being a crop of such importance there is naturally a great demand for accurate statistics as early as possible. The official forecast based upon plot-to-plot enumeration attempted to meet that demand. Sample surveys were also conducted by the Indian Statistical Institute. These were objective methods of enquiry where

⁴ For example, an assessment is sometimes made in regard to listing (of households for sample selection) by repeating the same sample areas (villages or urban blocks). A partial analysis of such data (NSS 13th round) shows that in two States (rural areas only) there has been under-listing of households by 0.7% and 1.0% respectively. Another example is provided by our sample survey of manufacturing industries where data on capital, labour, input, output and other related information are annually collected. In the current round data are being collected separately for the last two years. The data for the year before last were also collected in the previous round. As the larger establishments are being covered in both the rounds we have here an opportunity of examining the consistency of the information obtained in the two rounds for these common (about a thousand) establishments. A preliminary examination shows that this technique is likely to prove useful. Taxes of the order of 12.5 million rupees paid on account of excise duties have not been taken into account in the previous round. In another instance an amount of 5.6 million rupees was left out from the input entries in the earlier round.

EVALUATION OF STATISTICAL SURVEYS

a random sample of fields was taken up for actual physical examination for acreage estimation; and harvesting of a random sample of crop cuts followed by necessary weighments provided the yield rates. There were two interpenetrating sub-samples (IPNS) which were covered by different parties of investigators.

19. Table 1 gives the relevant information. It will be noticed that in both the years the official forecasts based on complete count were both very much out whereas the sample survey estimates were quite close to the trade figures. The two sub-sample (IPNS) estimates in both years agreed with the trade estimates within roughly 3 per cent while the estimates based on the so-called complete count differed from the trade figures by 27.2 per cent in 1944-45 and 16.6 per cent in 1945-46.

TABLE 1. COMPARISON OF OFFICIAL (COMPLETE COUNT) AND SAMPLE SURVEY
ESTIMATES OF JUTE PRODUCTION WITH TRADE FIGURES, BENGAL,
1944-45 AND 1945-46

particulars about jute crop	quantity (thousand bales) ^a	
	1944-45	1945-46
(1)	(2)	(3)
1. consumption during the season		
1.1 in jute mills (actual)	6000	6308
1.2 exports (actual)	1050	2213
1.3 in villages (estimate)	600	600
2. total	7650	9121
3. consumed from previous year's stock	324	697
4. jute crop in other provinces	598	862
5. balance : Bengal crop, trade figures	6728	7562
6. complete count : Bengal crop, official forecast	4895	6304
7. sample survey: Bengal crop, Indian Statistical Institute		
7.1 sub-sample 1 ^b	6836	7734
7.2 sub-sample 2 ^b	6518	7773
7.3 full sample	6686	7755
8. discrepancy of (6) on (5)	-27.2%	-16.6%
9. discrepancy of		
9.1 (7.1) on (5)	+ 1.6%	+ 2.3%
9.2 (7.2) on (5)	- 3.1%	+ 2.8%
9.3 (7.3) on (5)	- 0.6%	+ 2.6%

^a 1 bale = 400 pounds, 1 *maund* = 82.2857 pounds;

Trade figures are reported in bales but sample survey carried out on the basis of *maunds* which have been converted into pounds and then to bales, (the approximation 1 bale = 5 *maunds* sometimes used would give somewhat lower figures).

^b Sub-sample estimates of production have been obtained by multiplying sub-sample estimates of crop acreage by the full sample yield rates at district level, (IPNS estimates of yield rates being not easily available now).

III. SPOT CHECK OF CROP CENSUS RECORDS, 1937, 1949-50 AND 1950-51

20. For a very large portion of the cadastrally surveyed area of the country the statistics of acreage under important crop are obtained every crop season by the census method. The data are not collected by interviewing the cultivators. Actual visits to all the fields in a village are to be made; and therefore the data should

be relatively free from sizeable ascertainment errors, especially, when these are collected by the *patwari*, a village official, who is well informed about the crops grown in the locality.

21. The *patwari*, however, is a land revenue official and the crop records are primarily maintained for this purpose; and it is pertinent to enquire whether or not the data fulfil the more exacting requirements of crop statistics. He is moreover burdened heavily with multifarious duties and therefore actual visits to the fields may not always be possible.

22. A spot-check of the census records was therefore undertaken in the *Rabi* season, 1949-50, the work being under the control of the Department of Economic Affairs of the Ministry of Finance. The Government secured the services of a very senior officer with more than 20 years experience in land revenue work. Moreover, this officer had himself set up a crop census organisation in one of the States.

23. A batch of specially experienced investigators made visits to a sample of fields in the presence of the *patwaris*, and after securing their agreement noted the actual utilisation of the fields (whose areas were accurately known); the corresponding entries in the census records were also noted. A similar spot-check was again conducted during the *Rabi* season of 1950-51. A summary of the findings in respect of wheat, barley, gram, *arhar*, *mator*, mustard and linseed is given in Table 2.

TABLE 2. COMPARISON OF CROP ACREAGES AS REPORTED BY PATWARIS AND CHECKERS, 1949-50 AND 1950-51

name of crop ^a	number of villages	number of comparisons ^b	acreage		sum of		discrepancy as % of col. (5)	
			patwari	checker	positive discrepancies	negative discrepancies	absolute	algebraic
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. 1949-50								
1.1 wheat	391	8097	3492.0	3030.9	965.8	-504.7	49	15
1.2 barley	391	4144	1185.7	1117.8	468.5	-400.5	78	6
1.3 gram	391	6519	3010.3	3027.1	866.1	-883.0	58	-1
1.4 arhar	391	1120	298.9	494.1	57.9	-253.2	63	-40
1.5 mator	391	1340	195.7	260.4	54.2	-119.0	67	-25
1.6 mustard	391	3071	293.7	693.6	115.8	-515.8	91	-58
1.7 linseed	391	1145	134.0	269.2	59.3	-194.6	94	-50
2. 1950-51								
2.1 wheat	167	1170	1005.2	863.6	205.3	-63.7	31	16
2.2 barley	167	277	91.7	81.3	23.5	-13.1	45	13
2.3 gram	167	1819	1739.0	1301.0	602.8	-164.7	59	34
2.4 arhar	167	610	183.9	255.4	82.7	-154.2	93	-28
2.5 mator	167	357	105.5	189.9	15.6	-100.0	61	-44
2.6 mustard	167	247	58.1	38.0	26.3	-6.2	85	53
2.7 linseed	167	130	174.1	26.7	151.1	-3.6	580	552

^aCrops in mixture allocated to components in 1949-50, but completely excluded in 1950-51.

^bDoes not include where neither *patwari* nor checker reports the specified crop.

EVALUATION OF STATISTICAL SURVEYS

24. It should be pointed out at the outset that the 1949-50 results are not comparable with those of 1950-51 as mixed crops were excluded in the later year but not in the previous one. There are also differences in geographical coverage, and in the method of selection of sample villages and sample fields (plots). In the earlier check the choice of the villages was conditioned by the selection of *patwaris* made by their supervising *kanungos* and also by the avoidance of villages where harvesting had already started. Also in the selection of the plots within a sample village attempt was made to cast the sample haphazardly so as to cover all portions of the village. In the later spot-check both the villages and the fields were selected at random. However, a substantial number of these villages could not be covered by spot-check as the *patwari* could not be contacted, or even when contacted, he had not completed the crop records due to other preoccupations like the population or the livestock census. There were also some villages which did not fall under the jurisdiction of any *patwari*.

25. Growing of crops in mixture is fairly common during the *Rabi* season. In the case of fields with mixed crops, the proportion of the field under mixed crop was shown separately. Also the acreages under the different constituents were generally separately recorded on the basis of eye-estimation. However, for the 1949-50 data there were a few cases where this allocation was not recorded and the total area for these cases was divided equally among the constituent crops at the tabulation stage. For the 1950-51 data this question did not arise because the analysis was restricted to crops grown singly. The 1949-50 comparisons as given in the Table show large absolute discrepancies ranging from 49 per cent for wheat to 94 per cent for linseed. Part of the discrepancy is undoubtedly due to rather defective⁵ instructions to the *patwari*. For example, he is asked to ignore all minor constituents and report the entire area under the major component. The effect of this for a crop like linseed, which is known to be grown extensively as a minor component in a mixture, is obvious. Although as a component it may be minor its total contribution is very large compared to the acreage under pure (unmixed) linseed. The large negative algebraic discrepancy (-50%) for 1949-50 must be, at least partly, due to this factor. The position is similar with mustard. Some adjustments for the above shortcomings are reported to be made before the publication of official acreage statistics; obviously, the basis for such cannot be the conditions obtaining in the year for which the statistics are reported. The situation calls for drastic steps and not mere adjustments.

26. A second source of discrepancy for which all *patwaris* may not be fully responsible is the unavoidable adoption of a rule of allocation to components of mixed crop at the tabulation stage (1949-50). The checker was asked to record separately the acreages of the constituents of a mixed crop. But *patwaris* in certain States were, according to official instructions, not expected to show the separate allocations. In such States for purposes of official acreage statistics, allocation of

⁵ In relation to agricultural statistics, but presumably not so for land revenue purposes.

the area to different components is made at the level of district each comprising about 2000 villages on the average. This means that the total district area under a particular type of crop mixture is allocated to the components according to the ratio⁶ fixed long ago, usually at the time of last settlement of land revenue, such settlement operations taking place at intervals of 20 or 30 years. This is obviously unsatisfactory. (Incidentally the problem of completely satisfactory procedure of allocation of mixed crops is still awaiting solution).

27. The magnitude of the discrepancies observed in 1949-50, however, is so large that it is hardly possible to explain this to be solely due to causes for which the *patwari* is not directly responsible. To obtain a clearer grasp of the role played by the *patwari* the 1950-51 data were analysed only for plots (and sub-plots) for which crops in intermixture were not reported. For such comparisons the discrepancies must be attributable to *patwari* performance. There is unmistakable evidence that the census records were far from accurate; the absolute discrepancy ranged from 31 per cent for wheat to 580 per cent for linseed.

28. The 1949-50 spot-check was not on a random sampling basis, and the 1950-51 check, although planned on a random basis, could not be executed as such; and therefore it is strictly not possible to assess the *net* effect of the ascertainment error on the estimation of total acreage under a crop. The random components of the ascertainment errors of individual units may balance to some extent on aggregation. To study the extent of this balancing we are presenting below summary results of one of our older studies conducted in 1937.

29. A compact area, called *thana*, 78 square miles in area, (comprised of 10 'units' covering a total of 108 *mauzas* or villages) was completely covered by a census; the acreage under jute was recorded plot by plot, for more than one-sixth of a million fields. A sample of 11 *mauzas* was selected at random, and a second complete enumeration was independently carried out in these *mauzas* by a different set of investigators. Again a systematic sample of every 20th plot with a random starting point was independently enumerated in all the *mauzas*. (There were other types of re-enumeration which need not be stated here). The survey was conducted by the Director of Agriculture, Bengal, and the statistical analysis of the data was undertaken by the Indian Statistical Institute.

30. The average discrepancy between the jute acreages according to the two enumerations are given below for the different units of increasing average sizes, namely, plot (0.29 acre), village (462 acres), union (7.8 sq. miles) and *thana* (78 sq. miles). At the plot level the average discrepancy is about 54 per cent of the average jute acreage (according to census). The percentage discrepancy varied considerably from one sample village to another—the median value is 26 per cent and the five middle most values ranged from 12 per cent to 33 per cent. As far as can be judged from comparison of duplicated observations on an average of about 859 plots per

⁶ All these ratios were not (and still are not, in spite of best efforts) known to us at the time of analysis of data.

EVALUATION OF STATISTICAL SURVEYS

union the median percentage discrepancy at the union level comes to about 30 per cent; and for the *thana* the percentage discrepancy on the basis of 8586 plots (random systematic) works out to nearly 18 per cent.

31. From the magnitude of the discrepancies noted above it appears that a census may not provide accurate estimates for small administrative areas. The random component of the non-sampling error may balance considerably for very large administrative areas with the result that for these large areas the results may not differ appreciably from those obtained by a carefully conducted sample survey if the bias component is comparatively small. But care should be taken not to interpret that what holds good for a large area also holds good for each (or majority) of its constituent (small) administrative units.

IV. SAMPLE VERIFICATION OF LIVESTOCK CENSUS, 1956

32. In most of the States of the Indian Union, the eighth quinquennial livestock census was conducted during February to April 1956 and the data were subsequently brought up to the reference date of 15 April 1956. The enumeration was done by village officials like *patwaris*. The collection and compilation of these statistics were the responsibility of the State authorities. The Central Government in the Ministry of Food and Agriculture desired an independent sample verification by a different agency, and, as a consequence, the National Sample Survey (NSS) Organisation undertook this work during June-July 1956.

33. For rural areas 1624 sample villages, selected from among those just covered for socio-economic enquiries in the tenth round of NSS were taken up. Similarly a sample of 340 urban blocks (1951 Population Census Enumeration districts) were selected. For detailed enquiry 20 households were selected at random from each of the sample villages; and in urban blocks all the households were covered. In the sample units all non-household livestock establishments, which were very few in number, were also covered.

34. Data were collected about the livestock in possession of a household on the date of survey together with information about changes since 15 April 1956, so that the number as on the census date of reference could be obtained. The census registers were consulted only after the collection of these data, and census entries corresponding to the sample households were copied on the sample verification schedule. This was done in the sample village or urban area itself in order to minimise 'matching' difficulties.

35. It must be pointed out that the NSS investigators (enumerators) are wholtime quasi-permanent staff employed on a continuing basis, and are especially trained and experienced in the collection of data (including those on livestock) by the interview method, and their work is under intensive supervision. The data thrown up by them may therefore be considered to be of higher validity, and the census results may be evaluated against the results of the sample verification as standard.

36. Before passing on to the actual comparison of the results as shown in Table 3 we may refer to certain general features. Household by household enquiry revealed that in nearly one-fourth of the households having livestock the census enumerator did not make an actual visit. This fraction varied considerably from State to State. It did not however mean that the census enumerator made completely arbitrary entries for these households; he might have ascertained from neighbours, or, being a local man, the recording might even have been from his personal knowledge.

37. It was not always possible for the NSS investigators to secure the census figures for the sample households for various reasons. For example, the village might not have been covered in the census (at least up to the sample verification time), 'non-availability' of census records in the village (which may either mean that census had not taken place, or that for some reasons the records had been removed to some other place), 'clubbing' together two or more households in the census records, etc. Cases of omission to record the census entry against the verification figure might also have been caused by 'matching' difficulties, but such a contingency is likely to be very rare because the census enumerator (a local person usually the head-man or village *patwari*) from whom the census records were collected must have helped in identifying a sample household in his records unless, of course, when it was a case of omission on his part. In about 9.43 per cent of the rural and 5.61 per cent of urban sample households census figures were not recorded. A careful scrutiny of the investigators' 'remarks' showed that there were very strong grounds to conclude that some (at least one-eighth) of these cases were due to census omission.

38. We now pass on to the comparison of sample verification estimates with the census figures. We restrict ourselves to rural areas which account for by far the larger proportion of total livestock. Table 3 gives the necessary information not only for all-India but also for the major States. Manipur, Orissa and West Bengal are omitted because no census took place before the sample verification.

39. It will be noticed that there are highly significant differences in the all-India estimates of total cattle and total buffaloes. There is serious under-estimation in census figures, an increase of about 15 per cent is needed to bring up the census figure to the level of sample verification estimates. At the State level also there are significant differences.

40. Another set of estimates called the census-sample (cs) estimates has also been obtained in order to have some idea of the effect of different sources of error. Table 4 gives the necessary information for all-India (rural) in respect of total animals, total males, total females, working bullocks, and cows in milk, separately for cattle and buffaloes. The sample verification (sv), the census-sample (cs) and census (c) estimates are all shown in this Table.

41. In making the census-sample estimates we have made complementary use of census as well as survey data. From the census we have taken the counts of head of livestock for the sample households, and from the survey we have taken

TABLE 3. COMPARISON OF CENSUS COUNTS AND SAMPLE VERIFICATION ESTIMATES OF LIVESTOCK NUMBERS (MILLIONS): 1956

Rural sector: All-India and Major States

state ^a	sample surveyed			total cattle				total buffaloes				
	villages	h.hs.	census ^b	sample verifica- tion	difference	s.e.	diff. s.e.	census ^b	sample verifica- tion	difference	s.e.	diff. s.e.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
1. Andhradesh	88	1721	5.75	7.38	-1.63	0.64	2.55*	4.02	5.15	-1.13	0.42	2.69**
2. Assam	41	801	5.27	7.25	-1.98	0.81	2.44*	0.52	0.48	0.04	0.11	0.36
3. Bihar	192	3505	15.11	14.86	0.25	0.87	0.29	3.52	3.19	0.33	0.18	1.83
4. Bombay	119	2312	11.39	13.31	-1.92	0.80	2.40*	4.02	3.92	0.10	0.26	0.38
5. Hyderabad	80	1548	9.48	11.08	-1.60	0.76	2.11*	2.39	2.50	-0.11	0.16	0.69
6. Madhya Bharat	36	682	6.27	7.08	-0.81	0.49	1.65	1.91	2.39	-0.48	0.29	1.67
7. Madhya Pradesh	95	1866	15.35	15.83	-0.48	1.50	0.32	2.48	2.97	-0.49	0.39	1.26
8. Madras	139	2705	10.16	9.80	0.36	0.71	0.51	2.20	2.23	-0.03	0.22	0.14
9. Mysore	48	907	4.82	6.38	-1.56	0.65	2.40*	1.17	1.57	-0.40	0.16	2.50*
10. Punjab	55	1083	4.27	5.13	-0.86	0.52	1.65	2.76	3.29	-0.53	0.29	1.83
11. Rajasthan	66	1270	11.26	12.52	-1.36	0.82	1.54	3.20	3.59	-0.39	0.40	0.98
12. Uttar Pradesh	278	5526	22.52	27.92	-5.40	0.82	6.59**	9.57	11.47	-1.90	0.51	3.73*
13. All-India ^c	1434	27729	134.32	154.86	-20.54	2.97	6.92**	41.11	46.78	-5.67	1.11	5.11**

^aStates prior to 'Reorganisation'.^bAdjusted, whenever necessary, to States prior to Reorganisation on the basis of district census figures; magnitude of adjustment quite negligible compared to total.^cExcludes Manipur, Orissa and West Bengal, in which 4, 72 and 96 villages were respectively covered by NSS.

*Significant at the 5 per cent level.

**Significant at the 1 per cent level.

the counts of the number of households in the sample villages. In view of the fact that for some of the sample households census counts were not recorded for reasons explained earlier we proceeded as if those households were really taken into consideration by the census enumerator and assumed further that the average characteristics of those households were the same as those for which census data were available. (The latter assumption holds broadly for sample-verification data).

TABLE 4. COMPARISON BETWEEN SAMPLE VERIFICATION (sv), CENSUS-SAMPLE (cs) AND CENSUS (c) ESTIMATES : 1956

Rural sector : All-India^a

number of sample villages : 1334

number of sample households : 27729

livestock category	sample verification ^b (sb)	census sample ^b (cs)	census ^b (c)	adjustment factor		
				total sv/c	coverage-compilation cs/c	enumeration sv/cs
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. <i>cattle</i>						
1.1 total cattle	154.9	148.8	134.3	1.152	1.108	1.040
1.2 total males	86.2	83.2	73.8	1.168	1.126	1.037
1.3 total females	68.6	65.7	60.4	1.136	1.087	1.045
1.4 working bullocks	61.8	60.3	53.2	1.162	1.135	1.024
1.5 cows in milk	15.9	17.7	17.1	0.930	1.038	0.896
2. <i>buffaloes</i>						
2.1 total buffaloes	46.8	44.5	41.1	1.138	1.082	1.052
2.2 total males	12.0	11.2	10.5	1.136	1.068	1.064
2.3 total females	34.8	33.2	30.6	1.139	1.087	1.048
2.4 working bullocks	5.2	4.9	4.9	1.053	1.002	1.051
2.5 cows in milk	10.1	11.3	10.8	0.932	1.046	0.891

^a Excludes Manipur, Orissa and West Bengal.

^b Figures in millions.

42. The three estimates are subject to errors from various sources which may be broadly divided into the following classes: (1) coverage error—omission (or duplication) of villages or blocks, (2) listing error—omission (or wrong inclusion) of households in selected villages or blocks, (3) errors in enumeration (and categorisation) of livestock for the households concerned, (4) errors in compilation and editing etc. The sample-verification and census-sample estimates are in addition subject to sampling error. The listing error may be regarded as coverage error excepting in the situation described below. In this census there is reason to believe that in some instances the enumerators deliberately omitted to record households without livestock. To the extent this is done correctly the failure to list does not contribute anything towards coverage error. When, however, non-possession of livestock happened to be a wrong assumption such omission should strictly be regarded as enumeration error.

EVALUATION OF STATISTICAL SURVEYS

43. It is believed that sample-verification (sv) estimates are subject to comparatively small amount of errors of the above type. We may take the sample verification estimate as standard against which to compare the other estimates.

44. It is clear from the method of obtaining the census-sample (cs) estimates that these are at par with sample-verification (sv) estimates, excepting for errors of enumeration of livestock and therefore the ratio sample-verification estimate/census-sample estimate or sv/c may be taken as an index of (or an adjustment factor for) the net census-enumeration error. The last column in Table 5 gives estimates of enumeration adjustment factor.

45. Again one would expect agreement (excepting to the extent sampling error comes into play) between the census (c) estimates and census sample (cs) estimates if census figures were relatively free from coverage, compilation (including editing) errors. We can therefore use the ratio (census-sample estimate)/(census estimates) or cs/c as an index of (or adjustment factor for) *net* coverage-compilation error. The coverage-compilation adjustment factor ranges from 1.038 to 1.135 for the cattle categories noted in Table 4. The corresponding range for buffaloes is 1.002 to 1.087. Generally the coverage-compilation error affects the aggregate figures more seriously than enumeration errors.

46. An index or adjustment factor for the total error is given by the ratio sample-verification estimate/census estimate, or sv/c . It will be noticed that we have defined the adjustment factors in such a manner that total adjustment factor is equal to the product of the coverage-compilation adjustment and the enumeration adjustment factor.

47. Leaving apart cows in milk for the present the total adjustment factor is around 1.15 for the various cattle categories, and nearly 1.14 for the various buffalo categories (excepting working bullocks for which it is 1.05).

48. But for cows in milk, in spite of the fact that coverage-compilation adjustment factor is greater than unity, there is so much overenumeration in census that the total adjustment factor becomes less than unity. It is to be remembered, however, that placement in the category cows in milk is likely to be subject to considerable ascertainment error specially when the cow is nearing the end of its lactation period. Moreover, there is some amount of seasonal variation which might have added to the difficulty of ascertaining the position as on the census date; (both the census and the sample survey were spread over a fairly long period before and after this date respectively).

49. The total adjustment factors are themselves subject to errors of sampling. We have estimated these for the all-India categories, total cattle and total buffaloes,—the standard errors are 1.92 per cent and 2.36 per cent of the respective adjustment

TABLE 5. FREQUENCY DISTRIBUTION^a OF TOTAL (t), COVERAGE-COMPIATION^b (c), AND ENUMERATION^b (e) FACTORS
OF ADJUSTMENT : LIVESTOCK CENSUS AND SAMPLE VERIFICATION, 1956

adjustment factor	cattle															buffaloes														
	total cattle			total males			total females			working bullocks			cows in milk			total buffaloes			total males			total females			working bullocks			cows in milk		
	t	c	e	t	c	e	t	c	e	t	c	e	t	c	e	t	c	e	t	c	e	t	c	e	t	c	e	t	c	e
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)	(31)
1. 1.50-	—	—	—	1	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—
2. 1.40-1.50	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3. 1.30-1.40	2	1	—	2	3	—	—	1	—	2	2	—	1	1	—	1	—	—	4	1	—	—	1	—	1	—	—	—	1	2
4. 1.25-1.30	1	1	—	1	—	—	2	—	—	1	1	1	1	—	—	2	—	1	—	2	1	—	—	—	—	1	—	—	—	—
5. 1.20-1.25	3	—	1	1	—	1	4	1	2	1	1	—	1	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
6. 1.15-1.20	4	3	1	6	4	—	1	2	—	4	—	—	—	—	—	3	1	1	—	2	2	4	1	2	—	—	—	—	—	—
7. 1.10-1.15	2	4	—	1	2	1	4	3	1	1	3	1	1	2	—	2	1	2	3	2	2	2	—	—	—	—	—	—	—	—
8. 1.05-1.10	—	1	3	—	1	2	1	1	1	2	2	—	1	—	—	—	3	6	—	1	2	2	2	—	1	—	2	1	—	—
9. 1.00-1.05	1	1	4	1	2	6	1	2	5	2	1	6	2	3	2	4	—	2	1	—	5	—	—	1	4	—	2	4	—	—
10. 0.95-1.00	2	1	4	2	—	—	3	1	2	4	1	1	4	1	1	4	1	—	1	—	—	—	2	1	—	—	—	—	4	1
11. 0.90-0.95	—	—	—	—	—	—	—	—	—	—	—	—	1	2	—	1	2	1	3	1	1	1	3	1	1	1	1	1	3	2
12. 0.85-0.90	—	—	—	—	—	—	—	—	—	—	—	—	—	2	2	—	1	—	—	—	—	—	—	—	—	—	—	—	—	4
13. 0.80-0.85	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
14. 0.75-0.80	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
15. 0.70-0.75	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
16. 0.60-0.70	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
17. 0.50-0.60	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
18. —0.50	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
19. total	15	13	13	15	13	13	15	13	13	15	13	13	15	13	13	13	13	13	15	13	13	15	13	13	15	13	13	15	13	13

^aEach case corresponds to one of the following 14 States (rural) : Andhradesh, Assam, Bihar, Bombay, Hyderabad, Madhya Bharat, Madhya Pradesh, Madras, Mysore, Orissa, Punjab, Rajasthan, Uttar Pradesh and West Bengal, the 15th case being all-India (rural) which includes other (minor) States.

^bFor coverage-compilation and enumeration adjustment factors Orissa and West Bengal could not be included as Census took place after the sample verification.

EVALUATION OF STATISTICAL SURVEYS

TABLE 6. DISTRIBUTION OF SAMPLE HOUSEHOLDS BY THE NUMBER BY WORKING BULLOCKS (CATTLE) REPORTED
AS PER LIVESTOCK CENSUS AND SAMPLE VERIFICATION SURVEY : 1956

Rural sector : All-India
number of sample villages : 1434^a

working bullocks as per sample survey	working bullocks as per census																												not avail- able			
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)		(28)		
0	13110	199	260	8	13	3	5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1135	14733			
1	240	2308	261	19	12	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	327	3168			
2	337	196	5650	177	106	7	5	—	1	—	—	—	—	—	1	1	—	—	—	—	—	—	—	—	—	—	—	793	7275			
3	17	7	117	472	62	8	3	2	2	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	66	757			
4	23	4	102	52	849	33	27	2	2	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	99	1192			
5	—	—	10	2	31	91	17	1	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	6	159			
6	5	—	8	5	19	7	172	5	12	—	3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	13	249			
7	—	1	—	—	—	—	7	26	3	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	38	—		
8	1	—	1	—	5	1	14	2	44	2	3	—	3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	5	81	—		
9	—	—	1	—	1	1	1	1	1	7	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	15	—		
10	—	—	—	—	1	—	1	—	2	1	12	—	1	—	—	—	—	1	—	—	—	—	—	—	—	—	—	3	22	—		
11	—	—	—	—	—	—	—	—	—	—	1	5	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	7	—		
12	—	—	1	—	—	—	1	—	—	2	—	—	4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	8	—		
13	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1	—		
14	—	—	—	—	—	—	—	—	—	—	—	1	1	—	5	—	2	—	—	—	—	—	—	—	—	—	—	—	—	9	—	
15	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	2	—		
20	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	1	4	—	
21	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1	—	—	—	—	—	1	1	—	
22	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2	—	
24	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	1	—	
26	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	
27	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	1	—	—	
28	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	
30	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	
40	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—
total	13734	2716	6411	735	1099	152	253	37	68	10	25	6	10	1	5	2	4	1	3	2	1	1	1	1	1	1	1	1	2450	27729		

αExcludes Manipur, Orissa and West Bengal.

^aExcludes Manipur, Orissa and West Bengal.

factors. For enumeration adjustment factors the percentage standard errors are roughly half of those shown above.

50. All the adjustment factors have also been calculated for the rural areas of 12 major States. The census took place in Orissa and West Bengal subsequent to the sample verification. For these States it was possible to calculate only the total adjustment factor. We have presented in Table 5 a consolidated picture of the situation. Since the sample size for most of the States was comparatively small not much significance can be attached to individual adjustment factors. But taken together they reveal significant tendencies.

51. It will be noticed that except for cows in milk there is general tendency for adjustment factors to be greater than unity. This tendency is more marked in case of cattle. Although no great significance need be attached to specific cases one may conclude that there are in fact a few States where the actual number in certain categories exceeds the census number by 25 per cent or more.

52. In order to obtain a more detailed case by case picture of enumeration error we have obtained a few two-way tables showing the distribution of all-India (rural) sample households by number of head of livestock as reported by the NSS investigator and as recorded by the census enumerator. The table for total cattle would have been preferred but to save space we are presenting, by way of illustration of general features, Table 6 for working bullocks (cattle) only.

53. Out of nearly twenty-eight thousand households covered by NSS the comparison was possible for 91 per cent of cases. Out of these comparable cases in 38 per cent of the households there was no cattle according to NSS. But in nearly 5 per cent of these no-cattle households, the census recorded one or more head of cattle, the average being more than two. Among those with cattle, the census recorded complete absence in about 5 per cent of cases, and in two-thirds of former households there was complete agreement between the two agencies. In spite of this agreement the average discrepancy between the census and NSS entries was considerable in relation to the number of head of cattle possessed by a household. In fact, the standard deviation of discrepancy was as much as 56 per cent of the average number of head of cattle per household. The picture for working bullocks (cattle) is very similar, the corresponding standard deviation being of the same order. The gross enumeration-error was therefore very high.

54. We now come to two important points which are sometimes made in favour of census as against sample survey. First, the census can provide reliable counts for small administrative units (say, villages, each with 100 households on the average), and secondly, accurate measurement of change between two fairly close points of time are furnished by the census. It is obvious from the nature and magnitude of the enumeration and coverage compilation errors that the livestock census has hardly satisfied these objectives.

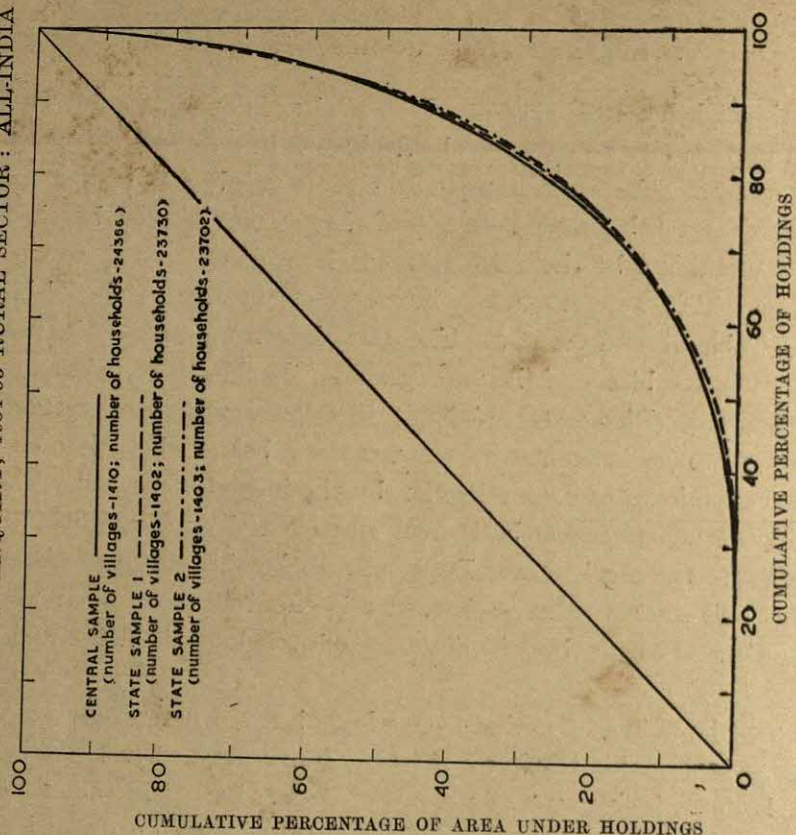
EVALUATION OF STATISTICAL SURVEYS

V. SURVEY OF LANDHOLDINGS, 1954-55

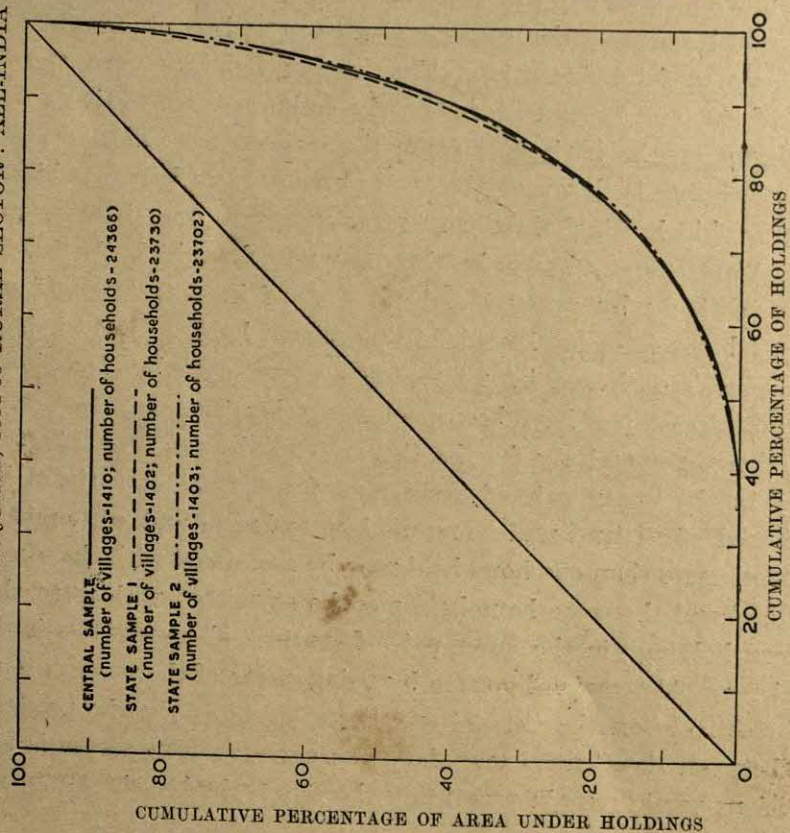
55. In the eighth round, 1954-55 of the National Sample Survey we had for the first time an opportunity to organise the field work in a completely independent interpenetrating system. The entire sample for a State was broken up generally into 12 independently drawn sub-samples, and 4 of these were covered by the Central agency and the remaining 8 by the State agency. In a few States there were additional sub-samples which were not taken into account in any of the Tables and Charts excepting Table 11 and Chart 3. The data thrown up opened the possibility of examining the relative biases of the two agencies. It should be pointed out here that the Central agency (Directorate of NSS) is a quasi-permanent agency purely for statistical surveys, the investigators (enumerators) being whole-time employees with considerable experience of data collection by the interview method. The State agencies participated on an *ad-hoc* basis, and employed their normal staff, usually of the Land Revenue and Agricultural Departments, for this purpose. An enquiry into landholdings and several other socio-economic enquiries were taken up simultaneously by the Central agency but the State agencies took up the first enquiry only.

56. It may be pointed out that even when observed differences are 'statistically significant', whether this is of any practical importance or not requires to be judged against the 'permissible error' for the purpose in view. The present enquiry, for example, was made primarily to collect information to decide broad policies of land redistribution; for such purposes the concentration curves (shown in Charts 1, 2, and 3) can supply very useful information. The sample holdings were arranged in ascending order of size and accumulated; and the estimated cumulative percentage of holdings is shown on the horizontal scale; the estimated cumulative percentage of the area under holdings is shown on the vertical scale. In Chart 1, which relates to 'household ownership holding' three concentration curves are shown separately, one for the Central sample, and two for State samples for which the information was collected respectively by party 1 and party 2 of investigators. Chart 2 is similar to Chart 1 but relates to 'household operational holding,' (which is defined as 'area owned' plus 'area leased in' minus 'area leased out' by household). In Chart 3, the two State samples have been pooled together; and only two concentration curves are shown respectively for the Central and the State sample for 'agricultural holding' in the upper pair of curve, and for '(total) operational holding' (which is constituted of all land under one distinct technical and economic unit so that in certain cases more than one household may be associated with the same 'total operational holding' and the same household may be associated with more than one '(total) operational holding') in the lower pair of curves. The real issue is whether the policy decision would remain the same by using *either* of the two concentration curves based respectively on the Central and State samples. In the present case, the divergence between the Central and the State concentration curves is so small that policy decisions would remain practically the same whether the Central or the

LANDHOLDINGS ENQUIRY, 1954-55 RURAL SECTOR : ALL-INDIA



LANDHOLDINGS ENQUIRY, 1954-55 RURAL SECTOR : ALL-INDIA



EVALUATION OF STATISTICAL SURVEYS

State concentration curve is used; the divergence between the two (Central and State) concentration curves may therefore be considered to be well within the limits of permissible error.⁷ It may be added that the converse case may also arise. The observed divergence between concentration curves or other results based on two sub-samples of an interpenetrating network of samples may not be statistically

LANDHOLDINGS ENQUIRY, 1954-55 RURAL SECTOR : ALL-INDIA

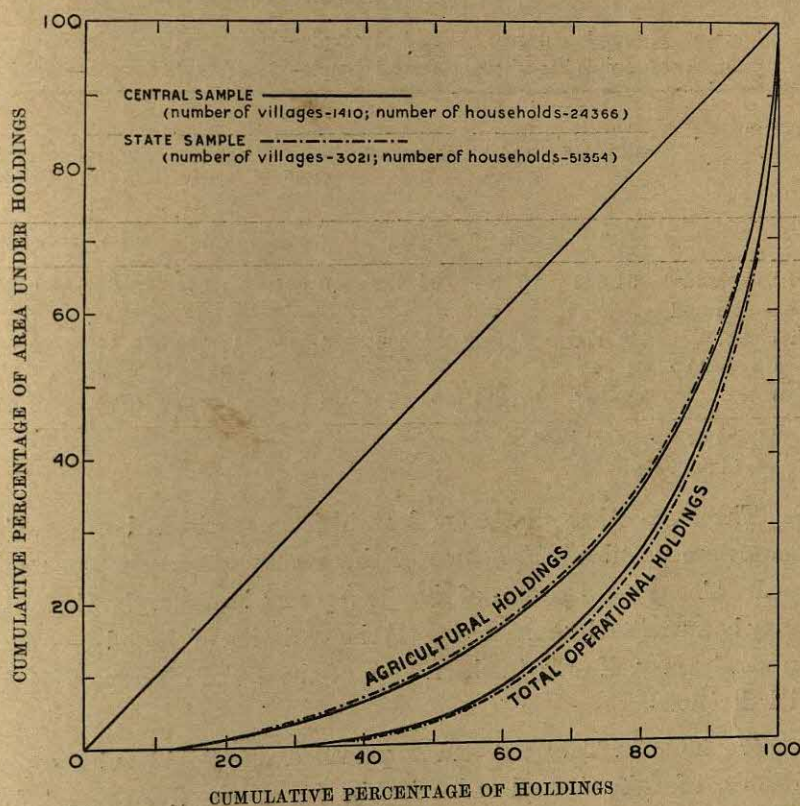


CHART (3) : CONCENTRATION CURVES FOR AGRICULTURAL HOLDINGS
AND (TOTAL) OPERATIONAL HOLDINGS

significant and yet be wider than the limits of 'permissible error' for any given purpose. The statistically 'significant' differences which will be found in some cases in the following paragraphs, while indicating some lack of control, should be judged, however, in the light of the observations made here.

57. We are giving in Table 7 the all-India (rural) estimates provided by the Central and State agencies for some basic characters. The corresponding standard errors are also shown. For every State separate estimates were obtained for each of the sub-samples; these facilitated the estimation of the standard errors. It will be noticed that the Central estimate is significantly larger than the State estimate in

⁷ The 'pooled' concentration curve which would lie 'between' the other two curves is not shown in any of the charts in order to avoid confusion arising from very close superimposition of several curves.

respect of 'number of households'⁸ as well as the 'acreage operated'. In case of 'area owned' although the former estimate is larger the difference is not statistically significant. We arrive at the same conclusion by a simpler method to be explained immediately.

TABLE 7. COMPARISON OF CENTRAL AND STATE ESTIMATES FOR SOME BASIC CHARACTERS : LANDHOLDINGS ENQUIRY, 1954-55

Rural sector : All-India

number of sample villages : central sample—1410; state sample—2805

number of sample households : central sample—24366; state sample—47432

character	central		state		difference		std. error of diff.(6) ^a	diff./ std. error
	estimate ^a	std. error ^a	estimate ^a	std. error ^a	actual ^a (4)-(2)	p.c. (6)/(2) %		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. number of households	64.863	0.848	62.892	0.504	1.971	3.0%	0.986	2.00*
2. total acreage operated								
by rural households	348.151	7.553	330.402	4.934	17.749	5.1%	9.022	1.97*
3. total acreage owned by								
rural households	307.037	7.148	304.992	4.780	2.045	0.7%	8.599	0.24

* Significant at 5 per cent level.

^a Figures in millions.

58. There are other characters for which comparisons between Central and State estimates are possible. For some of these comparisons simpler methods have been adopted. The general nature of the discrepancy between the Central and State estimates may be surmised by considering (1) the variation of the sub-sample estimates, and (2) the consistency, if any, in the nature of the set of discrepancies for a group of territorial divisions.

59. We have in Tables 8 and 9 provided all the 12 sub-sample all-India (rural) estimates, 4 for the Central and 8 for the States—for five classes of characters, namely, (1) aggregates, (2) average size, (3) rates per capita, in Table 8; and 4 distribution of households by size of holding, and (5) the area operated for each holding size class, in Table 9.

60. It may be pointed out incidentally that a few basic sub-sample estimates provide a simple means of obtaining sub-sample estimates of derived statistics, and thus the simple *t*-test which we have applied to the basic characters is also easily applied to the derived statistics. Thus we have given the results for the derived characters—total acreage leased in, the average sizes, and the rates per capita. Although not shown here, one can test out, from what is given, the significance of the difference in respect of average size for each holding size class.

⁸ The Central and State samples were drawn quite independently, with probability proportional to 1951 population, and there were a few, usually large, common villages, (villages were the first stage units); all the households in a village were first listed independently before the sample households were selected for detailed enquiry. A direct comparison of the exact number enumerated is therefore possible. The Central agency recorded on the average 11 households against 10 recorded by the State agencies.

TABLE 8. COMPARISON BETWEEN CENTRAL AND STATE SAMPLE ESTIMATES : AGGREGATES, AVERAGE SIZES AND RATES PER CAPITA : LANDHOLDINGS ENQUIRY, 1954-55

Rural sector : All-India^a

number of sample villages : sub-sample size—318

number of sample households : sub-sample size—5416

specification of estimate	aggregates				average size				area per capita			
	number (in millions)		area (million acres)		household operational holding (acres)		household operational holding (acres)		household operational holding (acres)		household operational holding (acres)	
	households	persons	operational holdings	operated	owned	leased in (net)	household (number of persons)	operational holding (acres)	household operational holding (acres)	household operational holding (acres)	operated (acres)	owned (acres)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
1. central estimates												
1.1. sub-sample 1	61.0	298.7	59.5	294.7	270.7	24.0	4.90	4.95	4.83	4.44	0.987	0.906
1.2. sub-sample 2	58.8	285.3	58.0	323.0	274.0	49.0	4.86	5.57	5.50	4.66	1.132	0.960
1.3. sub-sample 3	59.9	292.8	59.4	313.6	268.4	45.2	4.89	5.28	5.24	4.48	1.071	0.917
1.4. sub-sample 4	56.5	280.1	55.7	290.4	257.4	33.0	4.95	5.22	5.14	4.55	1.037	0.919
2. state estimates												
2.1. sub-sample 1	56.3	286.7	54.8	278.2	254.9	23.3	5.09	5.08	4.94	4.52	0.970	0.889
2.2. sub-sample 2	56.7	283.5	54.9	281.0	248.4	32.6	5.00	5.11	4.95	4.38	0.991	0.876
2.3. sub-sample 3	58.3	295.0	57.2	269.9	246.1	23.8	5.06	4.72	4.63	4.22	0.915	0.834
2.4. sub-sample 4	56.8	280.5	48.4	301.6	276.3	25.3	4.94	6.23	5.31	4.86	1.075	0.985
2.5. sub-sample 5	58.8	301.1	57.3	300.4	269.9	30.5	5.12	5.24	5.11	4.59	0.997	0.896
2.6. sub-sample 6	57.2	285.3	55.7	289.1	271.0	18.1	4.99	5.19	5.05	4.74	1.013	0.950
2.7. sub-sample 7	56.1	285.0	53.6	300.0	271.9	28.1	5.08	5.59	5.35	4.85	1.052	0.954
2.8. sub-sample 8	56.1	280.2	48.6	277.5	270.5	7.0	4.99	5.71	4.94	4.82	0.990	0.965
3. pooled estimates												
3.1. central	59.0	289.2	58.1	305.4	267.6	37.8	4.90	5.25	5.17	4.53	1.056	0.925
3.2. state	57.1	287.2	53.8	287.2	263.6	23.6	5.03	5.34	5.03	4.63	1.000	0.918
3.3. difference (3.2-3.1)	-1.98	-2.05	-4.33	-18.23	-4.02	-14.2	+0.13	+0.09	-0.14	+0.10	-0.056	-0.007
3.4. difference (%) (3.3/3.1)	-3.4	-0.7	-7.4	-6.0	-1.5	-37.6	+2.7	+1.7	-2.7	+2.2	-5.3	-0.8
3.5. <i>t</i>	2.43*	0.44	2.30*	2.23*	0.61	2.51*	3.79**	0.36	0.93	0.80	1.73	0.24

* Significant at 5 per cent level. ** Significant at 1 per cent level for *d.f.* = 10, *t*.₀₅ = 2.228, *t*.₀₁ = 3.169.^a Some States have been omitted for convenience; only the States listed in Table 10 are included.

TABLE 9. COMPARISON BETWEEN CENTRAL AND STATE SAMPLE ESTIMATES : DISTRIBUTION OF HOUSEHOLDS AND AREA OPERATED BY HOUSEHOLD OPERATIONAL HOLDING SIZE-CLASS : LANDHOLDINGS ENQUIRY, 1954-55

Rural sector : All-India

number of sample villages : sub-sample size—351

specification of estimate	number of households (million) in household operational holding size class (acres)					area operated (million) acres by household operational holding size class (acres)					number of sample households : sub-sample size—5983	
	number of sample households : sub-sample size—351					number of sample households : sub-sample size—5983						
	0.00	0.01-0.09	0.10-0.99	1.00-4.99	5.00-9.99	10.00 & above	0.01-0.09	0.10-0.99	1.00-4.99	5.00-9.99	10.00 & above	
1. central estimates												
1.1. sub-sample 1	4.08	14.12	8.72	19.95	10.39	8.93	0.36	4.0	53.8	72.5	201.8	
1.2. sub-sample 2	4.25	13.01	8.64	20.07	9.14	10.02	0.34	4.1	55.0	65.2	244.2	
1.3. sub-sample 3	3.68	16.05	7.29	20.17	9.63	8.93	0.43	3.5	54.1	67.2	229.6	
1.4. sub-sample 4	3.39	13.46	8.21	19.23	9.05	9.05	0.43	4.0	52.6	64.4	215.0	
2. state estimates												
2.1. sub-sample 1	8.05	10.58	7.71	17.60	9.18	8.64	0.30	3.7	47.4	64.7	204.2	
2.2. sub-sample 2	8.22	9.56	8.70	18.35	9.02	8.35	0.27	4.1	49.0	64.4	201.4	
2.3. sub-sample 3	8.29	10.49	9.66	18.65	8.82	8.77	0.33	4.5	50.8	62.5	196.3	
2.4. sub-sample 4	8.51	9.02	8.51	18.57	8.65	9.47	0.25	4.0	51.0	61.4	232.9	
2.5. sub-sample 5	9.17	9.85	7.95	20.44	8.63	8.89	0.29	3.8	54.4	62.1	220.1	
2.6. sub-sample 6	8.16	10.28	7.87	18.98	8.86	9.06	0.29	3.7	50.3	62.9	223.4	
2.7. sub-sample 7	8.29	9.06	6.84	10.36	8.55	9.30	0.29	3.3	50.4	61.1	218.9	
2.8. sub-sample 8	8.13	10.17	8.01	18.17	9.43	8.34	0.29	3.7	48.2	66.0	206.1	
3. pooled estimates												
3.1. central	3.89	14.15	8.21	19.85	9.58	9.18	0.39	3.90	53.9	67.3	222.7	
3.2. states	8.35	9.87	8.16	18.17	8.89	8.86	0.29	3.86	50.2	63.1	212.9	
3.3. difference (3.2-3.1)	+4.46	-4.28	-0.05	-1.08	-0.69	-0.32	-0.10	-0.04	-0.37	-0.42	-0.98	
3.4. difference(%) (3.3/3.1)	+114.7	-30.2	-0.6	-5.4	-7.2	-3.5	-25.0	-1.0	-6.9	-6.2	-4.4	
3.5. <i>t</i>	20.04**	7.82**	0.12	2.36*	2.56*	1.38	5.27**	0.20	3.26**	2.74*	1.09	

* Significant at 5 per cent level.

** Significant at 1 per cent level; for *d.f.* = 10, *t*._{.05} = 2.228, *t*._{.01} = 3.169.

EVALUATION OF STATISTICAL SURVEYS

61. It will be noticed from Tables 8 and 9 that the Central and State estimates are not always in agreement with each other. It is found that the State sample provides lower estimates of aggregates compared to the Central sample. There appears to be significant differences in regard to the estimation of 'total number of households', 'total number of operational holdings', 'total area operated' and 'total (net) area leased in' by all rural households from urban households. For these aggregates the State sample under-estimates by 3.4 per cent, 7.4 per cent 6.0 per cent and 37.6 per cent respectively.

62. It is of interest to note that the State sample provides a significantly lower estimate of 'total number of households', but the 'average household size' for that sample is significantly higher compared to the Central sample, and the product of the two, namely, 'total population' is not significantly different. It is difficult to interpret the result with confidence. One of the possibilities is that the State sample failed to cover fully the very small households, and the contribution of those small households to the total population size is small compared to the sampling fluctuations so that the above test failed to detect the discrepancy. Another possibility is that the State agencies in preparing the list of households in the sample villages leaned more on the list of cultivators which they had for purposes of land revenue, and moreover, they may not have always taken into account any recent partitioning of households which might really have taken place, if recording of such information did not make any material difference in regard to collection of land revenue.

63. It is only logical to compare the characteristics for each State separately, because agencies differed from State to State. It is to be noted, however, that the reduction of the sample size at the State level may make the detection of differences more difficult. It may on the other hand be pointed out that when pooled to all-India level the State bias (and errors) will balance to some extent and this may also make detection at all-India level difficult, even though sampling error may be reduced. We have therefore found it desirable to present a summary picture separately for the more important States also. For this purpose we have chosen the characteristic 'per capita area operated' the Central and State estimates of which do not differ significantly at the all-India level. We have also chosen for further examination the characteristics, 'number of households', 'number of operational holdings', 'total area operated' and 'average household size', for which the two estimates differ significantly.

64. It will be noticed in Table 10 that in majority of the States the sign of the difference between the Central and State estimates are the same as at the all-India level (shown in an earlier Table) for characteristics for which the difference is significant at the all-India level. Thus out of 14 comparisons the similarity holds in 10 cases for 'total number of households', 11 cases for 'total number of operational holdings', 12 cases for 'total area operated', and 10 cases for 'average household-size.' It is interesting to note that although the 'per capita area operated' estimates do not differ significantly at the all-India level, yet in as many as 13 cases

TABLE 10. PERCENTAGE DIFFERENCE BETWEEN CENTRAL AND STATE SAMPLE ESTIMATES : LANDHOLDINGS ENQUIRY, 1954-55
Rural sector : Principal States^a

state	sub-sample size		no. of h.h.s.		no. of opl. holdings		total area operated		average hh. size		per capita area operated	
(1)	(2)	(3)	diff. % ^b	t	diff. % ^b	t	diff. % ^b	t	diff. % ^b	t	diff. % ^b	t
1. Andhradesh	18	344	-0.54	0.19	-7.91	0.97	-3.78	0.34	+5.49	3.43**	-8.07	0.82
2. Assam	10	229	-8.00	0.86	-10.41	1.15	-33.09	2.45*	-1.86	0.22	-29.84	3.39**
3. Bihar	36	557	-0.39	0.06	-3.10	0.48	-4.02	0.59	+0.76	0.60	-4.01	0.79
4. Bombay	36	577	-1.29	0.43	-10.44	2.12	-4.09	0.66	-0.38	0.14	-2.47	0.49
5. Hyderabad	18	335	+7.26	1.23	+1.34	0.17	+24.08	2.29*	+4.59	1.67	+10.88	1.50
6. Jammu & Kashmir	19	229	+7.97	0.49	+8.92	0.54	+8.24	0.44	+5.12	3.06*	-4.52	0.53
7. Madhya Pradesh	27	403	-2.88	1.01	-11.39	1.41	-2.13	0.30	+1.11	0.35	-0.69	0.13
8. Madras	24	389	+0.19	0.07	+0.74	0.24	-1.78	0.23	-0.21	0.80	-1.93	0.26
9. Mysore	10	220	-8.96	2.20*	-12.12	3.06*	-8.13	1.13	+8.25	2.25*	-6.74	1.15
10. Orissa	18	377	-4.72	0.86	-3.57	0.71	-7.30	0.76	-0.22	0.08	-2.61	0.37
11. Punjab	15	275	-11.39	1.51	-26.90	2.15	-16.96	1.16	+9.20	2.73*	-11.03	1.05
12. Rajasthan	18	244	+1.10	0.28	-7.34	0.92	-13.18	1.11	+0.39	0.16	-13.83	1.48
13. Uttar Pradesh	51	873	-3.83	1.04	-5.97	1.53	-11.72	2.36*	-6.12	2.77*	-13.08	2.80*
14. West Bengal	18	364	-18.07	2.18	-18.85	2.23*	-19.04	2.72*	+5.33	2.04	-8.89	1.13

* Significant at 5 per cent level. ** Significant at 1 per cent level; for d.f. = 10, $t_{.05} = 2.228$, $t_{.01} = 3.169$.^a Only States with more than 200 households per sub-sample are included.^b Difference : (State estimate—Central estimate) expressed as percentage of Central estimate.

EVALUATION OF STATISTICAL SURVEYS

TABLE 11. PERCENTAGE DIFFERENCE BETWEEN CENTRAL AND STATE ESTIMATES IN REGARD TO AGRICULTURAL HOLDINGS; LANDHOLDINGS ENQUIRY, 1954-55

zone	number of sample villages		number of sample households		number of agricultural holdings			area under agricultural holdings			average size of agricultural holding			proportion of agricultural holdings to total operational holdings		
	central	state	central	state	central (000)	state (000)	% difference ^a	central (000 acres)	state (000 acres)	% difference ^a	central (acres)	state (acres)	% difference ^a	central	state	% difference ^a
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1. north	204	407	3484	6997	9237	8064	-12.70	44571	38997	-12.51	4.83	4.84	+0.21	.8026	.7451	-7.16
2. east	331	696	6402	12868	13865	11800	-14.89	59014	51383	-12.93	4.26	4.35	+2.11	.8048	.7517	-6.60
3. south	223	503	4255	9651	8802	8333	-5.33	43922	42452	-3.35	4.99	5.09	+2.00	.6650	.6431	-3.29
4. west	168	372	2820	5991	3936	3899	-0.94	49778	49578	-0.40	12.65	12.72	+0.55	.6696	.6789	+1.39
5. central	244	524	3771	8532	6807	6444	-5.33	85572	89394	+4.47	12.57	13.87	+10.34	.6916	.6674	-3.50
6. north-west	240	519	3634	7315	4768	4336	-9.06	63594	56423	-11.28	13.34	13.01	-2.47	.7714	.7180	-6.92
7. all-India	1410	3021	24366	51354	47415	42876	-9.57	346451	328227	-5.26	7.31	7.66	+4.79	.7423	.7039	-5.17

^a% difference : (State estimate—Central estimate) expressed as percentage of Central estimate.

the Central estimate exceeds the State estimates. This is, of course, a significant result.

65. Going back to the earlier Table 9 we note the sub-sample estimates of the distribution of households by the six size-groups of household operational holdings; (the sizes were obtained correct to two places of decimals, so that 0.00 means less than .005 acres)—the size classes being 0.00, 0.01–0.09, 0.10–0.99, 1.00–4.99, 5.00–9.99 and 10.00 & above. The corresponding acreages are also shown for each group (excepting for 0.00). It will be seen that the differences are significant for all the groups excepting 0.10–0.99 and 10.00 & above. There is considerable difference between the two estimates for the two lowest size classes, 0.00 and 0.01–0.09. (The State sample has registered a very large proportion in the 0.00 class). But the difference ceases to be significant when the two lowest size classes are merged together. It is difficult to offer any very convincing explanation of the phenomenon without further probing analysis. It is not known whether the State investigators were inclined to record more approximate figures (so that rounding off to 0.00 would be more frequent), or whether thinking that the main purpose of the enquiry is to collect information about *agricultural* holdings they have not paid adequate attention to recording areas under house-site.

66. Some of the total landholdings may be used exclusively for non-agricultural purposes. Excluding these we obtain holdings each of which is wholly or partly put to agricultural use. Zonal⁹ distribution of such agricultural holdings is shown in Table 11.

67. It will be noticed that relative to the Central sample the State sample under-estimates the number of such holdings (by 9.57 per cent). This feature is to be found in all the six zones. We have previously seen that taking total landholdings (which are comprised of all agricultural as well as non-agricultural lands) we have a similar phenomenon (under-estimation by 4.33 per cent) and it is pertinent to enquire whether or not the position regarding agricultural holdings is merely a reflection of the other phenomenon about total landholdings. It will be noticed from Table 11 that not only the State sample under-estimates the (total) operational holdings but even the proportion of agricultural holdings to (total) operational holdings is practically uniformly lower in the State sample. This may happen if the State sample behaves in the following manner—(1) larger omission of agricultural holdings, possibly the smaller ones, and/or (2) undue failure to record any agricultural utilisation, again, possibly in some of the smaller holdings. An examination of the data (not reproduced here) shows that in the lower size-classes the State sample registers a lower percentage of agricultural holdings. It will also be noticed in Table 11 that the average size of agricultural holding is greater in the State sample—possibly because of omissions of some of the smaller holdings.

68. The survey design described at the beginning of this section is incomplete in one respect. This is in regard to the assignment of different sub-samples to

⁹ In the 1951 Population Census the States were suitably grouped to form six Population Zones.

EVALUATION OF STATISTICAL SURVEYS

different parties of investigators. It will be recalled that there were four sub-samples in the Central sample; two of them were covered by the first Central party and the remaining by the second Central party. There were similarly two parties of State investigators. In our earlier discussions we have ignored the question of party differences.

TABLE 12. ANALYSIS OF VARIANCE OF AVERAGE AREA OWNED PER HOUSEHOLD;
LANDHOLDINGS ENQUIRY, 1954-55

state	sum of squares			mean square			F	
	agency d.f.=1	party d.f.=2	error d.f.=8	agency	party	error	agency	party
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. Uttar Pradesh	0.252	0.176	0.150	0.252	0.088	0.019	13.26**	4.63*
2. Bihar	0.004	0.246	0.919	0.004	0.123	0.115	28.75(r)	1.07
3. Orissa	0.001	0.381	1.864	0.001	0.190	0.233	233.00(r)	1.23(r)
4. West Bengal	0.583	0.249	0.998	0.583	0.124	0.125	4.66	1.01(r)
5. Assam	1.480	0.050	17.776	1.480	0.025	2.222	1.50(r)	88.88*(r)
6. Andhra	0.064	0.381	2.740	0.064	0.190	0.342	5.34(r)	1.80(r)
7. Madras	0.059	0.080	1.230	0.059	0.040	0.154	2.61(r)	3.85(r)
8. Mysore	0.254	0.0016	7.281	0.254	0.0008	0.910	3.58(r)	1137.5**(r)
9. Travancore & Cochin	0.441	0.631	2.395	0.441	0.316	0.299	1.47	1.06
10. Bombay	0.175	0.919	3.778	0.175	0.460	0.472	2.70(r)	1.03(r)
11. Saurashtra	15.072	5.941	71.547	15.072	2.970	8.943	1.69	3.01(r)
12. Madhya Pradesh	0.037	1.673	4.935	0.037	0.836	0.617	16.68(r)	1.35
13. Madhya Bharat	0.134	3.266	9.278	0.134	1.633	1.160	8.66(r)	1.41
14. Hyderabad	4.673	2.422	12.922	4.673	1.211	1.615	2.89	1.33(r)
15. Vindhya Pradesh	6.120	1.703	16.394	6.120	0.852	2.049	2.99	2.40(r)
16. Rajasthan	1.649	4.819	45.394	1.649	2.410	5.674	3.44(r)	2.35(r)
17. Punjab	1.397	3.945	9.973	1.397	1.972	1.247	1.12	1.58
18. PEPSU	0.771	4.990	25.181	0.771	2.495	3.148	4.08(r)	1.26(r)
19. Jammu & Kashmir	0.0004	0.083	1.334	0.0004	0.042	0.167	417.50*(r)	3.98(r)

$F_{.05}(1,8)=5.32$.

$F_{.01}(1,8)=11.26$.

$F_{.05}(8, 1)=238.9$.

$F_{.01}(8, 1)=5982$.

$F_{.05}(2,8)=4.46$.

$F_{.01}(2,8)=8.65$.

$F_{.05}(8, 2)=19.37$.

$F_{.01}(8, 2)=99.37$.

(r) indicates error/party or error/agency in F-ratio.

* Significant at 5 per cent level.

** Significant at 1 per cent level.

69. In order to test whether there is any significant difference between the party estimates (and between Central and State estimates) the technique of analysis of variance has been applied on the character 'area owned per household' as estimated from the survey. The results for each of the 19 States (having 12 replications or sub-samples each) are shown in Table 12. The total degrees of freedom (11) has been split up as follows—between agency (1), between parties within agency (2), and error (8).

70. It will be noticed that there was significant difference between parties within agency for the States of Uttar Pradesh, Assam and Mysore. Agency differences were significant for Uttar Pradesh and Jammu & Kashmir.

VI. SAMPLE SURVEY FOR ESTIMATION OF RATE OF GROWTH OF POPULATION
1958-59, 1959-60

71. It is not an uncommon practice in the NSS to have a built-in system of checking in the survey design. There is in the first instance the system of completely independent field work as well as processing—one by the Central agency and the other by State agencies. This interpenetrating arrangement provides two valid but independent estimates of all characters under study.

72. There are other systems which provide comparisons, the elements of which are all provided by the same agency. We shall illustrate one such method used in our current plan in regard to the estimation of rate of growth of population which is obviously of fundamental importance in any planning for national development. The survey, conducted by the interview method, is restricted to the rural areas in the first instance. Our vital registration system being rather unreliable we had to explore the possibilities of sample survey methods.

73. Control of sampling and other errors is sought to be achieved in two ways. First, control of sampling error by having a fairly large number of sample villages and canvassing necessary information from each and every household in them. It is also believed that the *prima facie* acceptability or non-acceptability of vital rates for the sample villages obtainable from their complete coverage has helped to reduce and control errors of omission etc. Again complete coverage has helped cross-checking of dates of births and deaths in the neighbouring households by studying the time sequence and the interval between those events; the deaths to single member households, no longer existing at the time of enquiry, are also obtainable only on the complete enumeration of villages by enquiring about such cases in, say, every 10th household. Also the village as a sample unit has definitely better technical and practical advantages over a sample of households in re-enumerating the population, narrated later. Secondly, by making the reference period sufficiently long and sampling fluctuations are reduced further. This step naturally increased the magnitude of ascertainment biases due mainly to recall lapse.¹⁰ A short reference period would increase the sampling error itself beyond desirable limits as the resources at our disposal did not permit us to increase the number of sample villages. Moreover, there were reasons like relatively larger border bias if too short a reference period was chosen.

74. By border bias we mean undue (net) inclusion or exclusion of vital events actually occurring at points of time around the end-points of the reference period. In the NSS the normal practice is to have as reference a period of specified length

¹⁰ It is the general practice in NSS to introduce in its earlier rounds some studies of a 'pilot' nature before taking up the subject in question in a full-fledged manner. These studies are not small scale *ad-hoc* ones, but are extensive surveys so that the effect of large-scale operational conditions may be adequately reflected in the results. For similar reasons the exploratory nature is not usually fully articulated. The experience gained will help in reducing errors and biases in subsequent surveys. Earlier studies in respect of collection of data on vital events showed that under-reporting increased considerably with increase in recall period.

EVALUATION OF STATISTICAL SURVEYS

immediately preceding the date of interview, and such border bias is likely to be more serious near the remoter end-point.

75. It has been planned to collect data with a 2-year reference period and have such additional information on the time of actual occurrence that one can work with any shorter reference period at the tabulation stage. It is felt that although there would still be border errors for any smaller reference period yet it will mainly be of a random nature with the result that the net effect (bias) would be smaller than what it would be if the reference period of collection was identical with reference period of analysis. Conscious efforts made by the investigator and/or informant for the correct placement of a vital event in relation to the (remoter) cut-off point is believed to result in larger biases than what would result from a cut-off at the tabulation stage. It should be pointed out here that the rural population has usually no precise knowledge of the exact date of occurrence of a birth or death.

76. After establishing all these precautions it is necessary to have a self-evaluating system. In this connection we shall take up several points one by one. NSS investigators have acquired some experience in collection of birth and death data in earlier rounds. But it is not known whether they have reached a 'steady' state so that there is no further learning-effect. Evaluation in this regard is necessary. If the learning-effect is still found to be present then caution should be exercised in interpreting the results. For this and other purposes the NSS investigation is spread uniformly over one-year round. To be more specific the entire round has been broken up into six two-months sub-rounds and data have been separately tabulated for each sub-round. Sub-round comparisons will be of help in studying the learning-effect.

77. It is considered desirable for the sake of accuracy to enumerate each and every individual in all the households in the sample villages. This will help us to have not only a correct picture of the number in different age-sex-marital status classes but would also help us to record more correctly all births and deaths occurring to a member of the household canvassed. But as a multi-purpose survey organisation the NSS has to explore the effect of adopting a less time-consuming plan so that resources for other enquiries may be found. It has therefore been decided that in the 14th round only the first two sub-rounds would have a detailed individual by individual enumeration, whereas in the later sub-rounds summary information on a household basis would be collected. A comparison of the results thrown up by the two methods of varying intensity will therefore be of value in assessing the results thrown up by the less intensive enquiry.

78. There is also the basic problem of evaluation of the very approach of estimating birth and death rates as described earlier. A second more direct approach, particularly in regard to deaths, may be helpful in studying the efficacy of this approach. In the second approach an account has been taken of the whereabouts, in the fifth and sixth sub-rounds of the 14th round, of each and every individual

enumerated in the first two sub-rounds,¹¹ and this resulted in recording deaths if the person in question had died. Cases of migration, both in and out, have been also noted. Additions due to births and cases of omissions in either sub-rounds are also recorded. It may be pointed out that babies born immediately before the date of interview, even if omitted in the first two sub-rounds, (which may not be a very unlikely event), are less likely to be omitted at the re-enumeration stage when they have somewhat grown up (provided they have not died in the intervening period).

79. At the re-enumeration stage an attempt has also been made to collect data about births and deaths during the reference period. It will be seen from what has been said about that the second approach (in which, strictly speaking, the first approach is a part) is expected to be less liable to omission of deaths and to some extent of births as well. But both of them are likely to suffer from under-enumeration of cases of births and deaths of very short-lived infants. However, it may be pointed out that as our principal objective is to determine the rate of increase such cases would not affect (excepting in a marginal way) the results.

80. It must be emphasised, however, that exploring the possibility of estimation of birth and death rates fairly accurately has been also kept in view. It therefore becomes necessary to provide for a method of assessing the magnitude of the recall biases. Tabulation of data by varying lengths of recall periods will throw some light, but in addition a more direct approach appears desirable.

81. The evaluation plan takes into account not a single annual round but two consecutive rounds (the 14th and 15th rounds of NSS). According to this plan the same set of villages are being covered in these rounds. It will be recalled that our reference period is 'last two-years' so that, what is 'last year' in the earlier round becomes 'year before last' for the next round. We shall therefore have two sets of data for the same reference period, but with different recall periods. The 15th round is currently going on and it is hoped that the two sets will provide a means for evaluation of the accuracy of such data collected by the interview method. It will also provide a method of adjustment of data subject to recall lapse.

82. We shall now present some of the preliminary results obtained so far. Table 13 sets down the estimates of (1) average household size, (2) sex-ratio, (3) percentages of the population falling in the three age groups : 0-14, 15-44, 45-, (4) birth-rate, (5) death-rate and (6) rate of natural increase for each of the six sub-rounds of the 14th round (the first round for the 2-round enquiry). There are two parties of investigators; in every stratum one set of six villages is to be covered by an investigator belonging to the first party, and a second set of six by an investigator belonging to the second party. It is thus possible to obtain two valid and independent sets

¹¹ To ensure careful re-enumeration work, certain fictitious names have been entered in the original listing. And a new item of information viz., 'days sick during last month' has to be entered (in the fifth and sixth sub-rounds), so that on the one hand the investigator has to ask pointed questions about an individual earlier enumerated, and this would on the other hand incidentally help him to secure indirectly the delicate information about death which may have taken place.

EVALUATION OF STATISTICAL SURVEYS

of estimates. These are also shown. The reference period is 'last year' for births and deaths.

83. It appears on examination of all possible long *versus* short schedule comparisons that the shorter schedule gives rise to larger estimates of average household size. In regard to sex-ratio, there is no obvious difference. For the age groups, particularly the two older groups 15-44 and 45-, there is clear evidence of difference in the two schedule types. Under-estimation of the older groups and over-estimation of the middle group in the short schedule is obvious. For the younger group, 0-14, within party (or within sub-sample) comparisons show that the short schedule gives underestimates; 14 out of 16 possible comparisons support this.

TABLE 13. COMPARISON OVER SUB-ROUNDS, SCHEDULE TYPES AND SUB-SAMPLES OF CERTAIN DEMOGRAPHIC VARIABLES; POPULATION ENQUIRY, 1958-59

Rural sector : All-India										
sample size : 218 villages/sub-sample/sub-round										
2616 villages during entire round										
sub-round	schedule type	sub-sample or party	average household size	sex ratio ^a	percentage in age-group			rates per 1000 persons		
					0-14	15-44	45-	birth	death	growth
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1	long	1	5.07	104.12	40.70	42.84	16.46	38.37	18.98	19.39
		2	5.01	103.97	40.54	42.79	16.67	38.56	19.12	19.44
		pooled	5.04	104.05	40.62	42.82	16.56	38.47	19.05	19.42
2	long	1	4.97	101.50	40.90	42.65	16.45	38.76	19.50	19.26
		2	5.11	103.86	41.03	42.77	16.20	40.45	19.89	20.56
		pooled	5.04	102.74	40.97	42.72	16.31	39.66	19.70	19.96
3	short	1	5.16	102.61	40.45	45.16	14.39	37.49	19.82	17.67
		2	5.15	103.84	40.13	45.36	14.51	35.00	18.81	16.19
		pooled	5.16	103.17	40.31	45.25	14.44	36.36	19.36	17.00
4	short	1	5.21	104.94	40.65	45.36	13.99	41.63	21.57	20.06
		2	5.21	105.47	40.83	45.01	14.16	38.65	19.08	19.57
		pooled	5.21	105.20	40.74	45.19	14.07	40.13	20.32	19.81
5	short	1	5.14	103.13	40.62	44.60	14.78	36.33	20.04	16.29
		2	5.15	103.59	40.33	44.70	14.97	37.33	17.24	20.09
		pooled	5.15	103.37	40.46	44.66	14.88	36.85	18.60	18.25
6	short	1	5.28	103.60	40.83	45.13	14.04	38.13	16.28	21.85
		2	5.16	103.61	40.29	45.18	14.53	37.65	17.22	20.43
		pooled	5.22	103.61	40.53	45.16	14.31	37.87	16.79	21.08
1-6	long and short	1	5.14	103.30	40.68	44.31	15.01	38.50	19.47	19.03
		2	5.13	104.07	40.54	44.28	15.18	38.02	18.58	19.44
		pooled	5.14	103.69	40.61	44.29	15.10	38.26	19.02	19.24

^a Males per 100 females.

84. Similar analysis of birth and death rates does not reveal any marked difference between the two schedule-types excepting possibly for birth-rate which is higher (13 out of 16 possible comparisons) for the long schedule.

85. It is also necessary to study the learning or familiarity effect. The difference, if appreciable, will be revealed by an examination of the differences between sub-rounds other conditions remaining the same. We shall study this in respect of birth rates. Concentrating on the first two sub-rounds we note that both the parties have registered a higher rate at the second sub-round showing perhaps some improvement in enumeration of births. To secure a little more confirmation we have examined this in relation to the five (administrative) Zones into which the country has been divided; 8 out of 10 possible within-party comparisons support this. It will be recalled that from the third sub-round onwards the short schedule has been used. The fourth sub-round again shows a higher birth-rate compared to the third not only for all India but for 9 out of 10 possible within-party zonal comparisons. In case of death-rate, however, there is no such phenomenon. It would appear therefore that there is a learning-effect in regard to births but nothing very much in respect of deaths.

86. The existence of learning-effect brings in the question whether the difference in the schedule types could not be explained by the learning-effect. We would be inclined to feel that the nature and magnitude of the differences are such that it cannot be explained in that way.

87. Preliminary analysis of the re-enumeration data collected in the fifth and sixth sub-rounds show that the annual (the interval between enumeration and reenumeration was roughly 8 months¹² (number of deaths etc., has been proportionately increased to 12 months) death rate among persons enumerated earlier comes out 1.71%. This may be seen against a death rate of 1.77% obtained by the previous approach for the part of the enquiry conducted during the same survey-period (fifth and sixth sub-rounds). It will be noted that inclusion of infants born after the first enumeration and the deaths among them (before re-enumeration) would have made the two figures comparable. As far as one can make out from our estimate of birth rate viz., 3.82%, and the infant death proportion of 99 per 1000 live births (proportion of infants born alive and dying in the reference year) estimated from this enquiry, the comparable death-rate comes to about 1.95%. It would appear therefore that the method of accounting in re-enumeration has resulted in a more complete enumeration of deaths. Further probing analysis is necessary before one can be on surer grounds regarding this evaluation of the results.

88. It is necessary to examine the magnitude of recall-bias. A direct evaluation of this bias will be possible, as explained earlier when the 15th round data (yet to be collected and analysed) are seen against the 14th round data for the same villages and households. We can however study the effect in a less direct way from the 14th round data. It is found that at the all-India level, the 'year before last' birth rate was 81.7% of the 'last year' birth rate. In regard to death rate the corresponding percentage was substantially lower, viz., 52.6%. While the effect of the

¹² An eight month gap is not entirely satisfactory, one year would be preferable, and the data currently being obtained in the 15th round will remove this defect.

EVALUATION OF STATISTICAL SURVEYS

recall-lapse is quite appreciable in case of the above two rates it is interesting to estimate the effect of this on the growth-rate, because there is a certain amount of compensation as both are under-estimates and the one (death rate) subject to higher (%) recall-lapse has a lower absolute value. The net effect is over-estimation of growth rate by 10.5% if 'year before last' replaces 'last year'.

89. The above however is a partial appraisal of the situation because the question of recall-lapse within the course of the year has been overlooked. One can, however, reasonably expect from what has been said above that within year recall-lapse effect on the growth-rate cannot be very large. An actual study on the 7th round data showed that a slight increase of about 2 per cent in the growth rate may perhaps be allowed for recall-lapse within the course of the one year reference period.

90. It would appear therefore that the 14th round NSS estimates of growth-rate is perhaps subject to a slight downward bias. It is however not quite clear at this stage about the effect of a 2-year reference period, introduced for the first time in the 14th round, and how far the conditions obtaining in the 7th round are obtainable in the 14th round. Deeper analysis of 14th and 15th round material may help in clearing up the issues.

VII. ACKNOWLEDGEMENTS

91. In the preparation of this paper some of our colleagues have helped us with necessary data. We should like to acknowledge our obligation to all of them and particularly to A. Ganguly, R. P. Saha, R. K. Som and S. Raja Rao.

RÉSUMÉ

En développant les techniques pour l'évaluation des résultats d'inspection, l'on considère le fait suivant avec une importance qu'il mérite. Le résultat final dépend non seulement de la "vraie" valeur de l'unité pour l'observation, mais aussi de la méthode d'obtenir de telles renseignements, unité par unité, et aussi de la méthode d'arranger des données.

Le résultat peut même dépendre de la choix de l'unité de certitude ("ascertainment unit"). La technique fondamentale est donc l'établissement des comparaisons entre deux (ou un plus haut nombre) estimations de la même quantité obtenue sous les conditions opératoires et/ou techniques, partiellement ou entièrement différentes.

L'on peut évaluer l'inspection contre les données obtenues par une méthode qui est, a priori, plus valable. "Unitary checks" qu'il est possible d'introduire quand les unités de certitude sont identiques, mettent en jour les individuelles erreurs de certitude. La comparaison entre "totaux" évalués—et c'est possible quand même les unités de certitude sont différentes—donne l'erreur totale dérivée de tous les sources. L'évaluation des unités choisies aléatoirement nous fournit un moyen de juger de l'effet de l'opportunité amoindrie de compenser les erreurs émanées des sources différents, où il s'agit des estimations de 'totaux' ou quelque semblable sommaire de "renseignements" des unités territoriales (plus petites).

Cette évaluation est importante, car c'est possible que la petitesse de l'erreur totale quand les unités territoriales sont plus grosses, n'indique la même chose pour les petites unités. L'on développe une technique pour séparer la certitude totale ou l'erreur d'énumération, de l'erreur totale de "compilation-editing," en évaluant un "total" ou l'information sommaire semblable.

Dans le cas d'échantillons entrecroisés, quand l'on tire deux (ou un plus haut nombre) échantillons de la même population et si l'on les traite par le même dessein d'inspection, les résultats basés sur les différents échantillons sont également valables,—quand même ils sont obtenus par différents chercheurs et/ou par les différentes unités d'arrangement. Les différentielles opératoires correspondent à la divergence entre différentes estimations. De plus, l'estimation de l'étendue d'incertitude est basée sur cette divergence.

Une technique de la "propre évaluation" (self-evaluation) paraît aussi d'être utile. Pour achever ce but il faut accumuler les données sur les unités de certitude ayant différents "dimensions",—les unes étant les parties des autres—par exemple, une petite coupe dans une plus grosse coupe (des expériences de faucher). Les données pour une longue période de référence, accumulées en une telle manière que les réductions en tables pour une plus courte période de référence sont aussi possible, donnent, en effet, des différents dimensions de la période de référence. Une comparaison des estimations que l'on puisse obtenir des différents dimensions facilite l'évaluation.

La comparaison d'estimation, sur différents points ou périodes (de temps) est une autre méthode. Les comparaisons des estimations pour la même période de référence (mais aux différentes périodes d'inspection) sont aussi avantageuses. Faire des recherches sur l'accord dans les propriétés de la divergence entre différentes sections territoriales, en vaut la peine.

Toutes les techniques discutées sont nées des résultats obtenus dans les domaines variés—moisson, bétail, propriété et population.

ON CONFIDENCE INTERVAL FOR TWO-MEANS PROBLEM BASED ON SEPARATE ESTIMATES OF VARIANCES AND TABULATED VALUES OF t -TABLE

By SAIBAL BANERJEE

Indian Statistical Institute

SUMMARY. Given k samples of n_i units from k normal populations $N_i(m_i, \sigma_i^2)$ ($i=1, 2, \dots, k$) a confidence interval for any linear function $\sum_1^k c_i m_i$ (where c_i are known coefficients) with confidence coefficient not less than some pre-assigned probability α is possible in terms of sample estimates of population means and variances and tabulated values of Student's t -table. Some generalizations of the result including testing of hypothesis have been considered.

1. INTRODUCTION

Given two samples of n_i ($i = 1, 2$) units from two normal populations with means m_1 and m_2 and variances σ_1^2 and σ_2^2 a confidence interval for any linear function $c_1 m_1 + c_2 m_2$ (where c_1 and c_2 are known coefficients) in terms of sample estimates of population means and variances is of interest. It was shown by the author (Banerjee, 1960) that given k -samples of n_i units from k normal populations $N_i(m_i, \sigma_i^2)$ ($i = 1, 2, \dots, k$) and some pre-assigned probability α , a confidence interval for $\sum_1^k c_i m_i$ with confidence coefficient not less than pre-assigned probability α , could be built up from the relation

$$\text{prob} \left[\left\{ \sum_1^k c_i (\bar{x}_i - m_i) \right\}^2 \leq \sum_1^k \frac{t_i^2 c_i^2 s_i^2}{n_i} \right] \geq \alpha \quad \dots (1.1)$$

where $\bar{x}_i$ and s_i^2 ($i = 1, 2, \dots, k$) are sample estimates of population means and variances, c_i ($i = 1, 2, \dots, k$) are known coefficients and t_i ($i = 1, 2, \dots, k$) are so chosen that

$$\frac{1}{\sqrt{v_i}} \frac{1}{B\left(\frac{v_i}{2}, \frac{1}{2}\right)} \int_{-t_i}^{t_i} \left(1 + \frac{t^2}{v_i}\right)^{-\frac{v_i+1}{2}} dt = \alpha \quad (v_i = n_i - 1; i = 1, 2, \dots, k).$$

The probability statement (1.1) or the confidence interval associated with the probability statement (1.1) is applicable to more general situations. Some generalizations are considered in this paper.

2. NOTATIONS

The following notations have been used throughout this paper:

$f(\chi_i^2)$ denotes the frequency function of a χ_i^2 variate which is distributed as a χ^2 variate with v_i degrees of freedom.

$h(\chi^2)$ denotes the frequency function of a χ^2 variate which is distributed as a χ^2 variate with one degree of freedom.

$f(t/\nu_i)$ denotes the frequency function of a t -variate which is distributed as Student's t -variate with ν_i degrees of freedom.

The terminology that t_i ($i = 1, 2, \dots, k$) are t -values of Student's t -table of ν_i ($i = 1, 2, \dots, k$) d.f. corresponding to confidence coefficient α denotes t -values of Student's t -table so that

$$\int_{-t_i}^{t_i} f(t/\nu_i) dt = \alpha.$$

3. THEOREMS

Theorem 1 : If u be a normal variate distributed about zero mean and unit variance and χ_i^2 ($i = 1, 2, \dots, k$) be χ^2 variates distributed mutually independently and also independently of u with ν_i ($i = 1, 2, \dots, k$) degrees of freedom and w_i ($i = 1, 2, \dots, k$) be a set of arbitrary weights satisfying the relation

$$\sum_1^k w_i = 1; \quad w_i \geq 0; \quad (i = 1, 2, \dots, k)$$

then
$$\text{prob} \left\{ u^2 \leq \sum_1^k \frac{t_i^2}{\nu_i} w_i \chi_i^2 \right\} \geq \alpha \quad \dots (3.1)$$

where t_i ($i = 1, 2, \dots, k$) are tabulated values of Student's t -table of ν_i ($i = 1, 2, \dots, k$) degrees of freedom corresponding to confidence coefficient α .

Proof : We have

$$\text{prob} \left[u^2 \leq \sum_1^k \frac{t_i^2}{\nu_i} w_i \chi_i^2 \right] = \int_0^\infty \int_0^\infty \dots \int_0^\infty \prod_1^k f(\chi_i^2) \left\{ \int_0^R h(\chi^2) d\chi^2 \right\} d\chi_1^2 \dots d\chi_k^2. \quad \dots (3.2)$$

where

$$R = \sum_1^k \frac{t_i^2}{\nu_i} w_i \chi_i^2,$$

As $\int_0^z h(\chi^2) d\chi^2$ is an upward convex function of z ,

$$\sum_1^k \frac{t_i^2}{\nu_i} w_i \chi_i^2 \geq \sum_1^k w_i \int_0^{\frac{t_i^2}{\nu_i} \chi_i^2} h(\chi^2) d\chi^2. \quad \dots (3.3)$$

It can be shown

$$\int_0^\infty f(\chi_i^2) \left\{ \int_0^{\frac{t_i^2}{\nu_i} \chi_i^2} h(\chi^2) d\chi^2 \right\} d\chi_i^2 = \alpha; \quad (i = 1, 2, \dots, k). \quad \dots (3.4)$$

CONFIDENCE INTERVAL FOR TWO-MEANS PROBLEM

As $\sum_1^k w_i = 1$, from (3.2), (3.3) and (3.4),

$$\text{prob} \left\{ u^2 \leq \sum_1^k \frac{t_i^2}{v_i} w_i \chi_i^2 \right\} \geq \alpha. \quad \dots (3.5)$$

Theorem 2 : Let Y be a normal variate distributed about mean M with variance

$\sum_1^k \lambda_i \sigma_i^2 + \sum_1^l \theta_j \sigma_j^2$, where λ_i and θ_j ($i = 1, 2, \dots, k; j = 1, 2, \dots, l$) are known positive constants. If σ_j^2 ($j = 1, 2, \dots, l$) be known and s_i^2 be estimates of σ_i^2 ($i = 1, 2, \dots, k$) where $v_i s_i^2 / \sigma_i^2$ ($i = 1, 2, \dots, k$) are mutually independently (and also independently of Y) distributed as χ^2 with v_i ($i = 1, 2, \dots, k$) degrees of freedom, then

$$\text{prob} \left[(Y-M)^2 \leq \sum_1^k t_i^2 \lambda_i s_i^2 + d^2 \sum_1^l \theta_j \sigma_j^2 \right] \geq \alpha \quad \dots (3.6)$$

where t_i ($i = 1, 2, \dots, k$) and d are respectively tabulated values of Student's t -table of v_i ($i = 1, 2, \dots, k$) d.f. and the normal probability table corresponding to confidence coefficient α .

Proof: We have $\text{prob} \left[(Y-M)^2 \leq \sum_1^k t_i^2 \lambda_i s_i^2 + d^2 \sum_1^l \theta_j \sigma_j^2 \right]$

$$= \int_0^\infty \int_0^\infty \dots \int_0^\infty \prod_1^k f(\chi_i^2) \left\{ \int_0^R h(\chi^2) d\chi^2 \right\} d\chi_1^2 \dots d\chi_k^2 \quad \dots (3.7)$$

where
$$\chi^2 = (Y-M)^2 / \left(\sum_1^k \lambda_i \sigma_i^2 + \sum_1^l \theta_j \sigma_j^2 \right)$$

$$\chi_i^2 = \frac{v_i s_i^2}{\sigma_i^2} \quad (i=1, 2, \dots, k)$$

$$w_i = \frac{\lambda_i \sigma_i^2}{\sum \lambda_i \sigma_i^2 + \sum \theta_j \sigma_j^2} \quad (i = 1, 2, \dots, k)$$

$$w'_j = \frac{\theta_j \sigma_j^2}{\sum \lambda_i \sigma_i^2 + \sum \theta_j \sigma_j^2} \quad (j = 1, 2, \dots, l)$$

$$R = \sum_1^k \frac{t_i^2}{v_i} w_i \chi_i^2 + d^2 \sum_1^l w'_j$$

and $f(\chi_i^2)$ denotes the frequency function of a χ_i^2 variate which is distributed as a χ^2 variate with v_i degrees of freedom and $h(\chi^2)$ denotes the frequency function of a χ^2 variate with one degree of freedom.

As $\int_0^z h(\chi^2) d\chi^2$ is an upward convex function of z ,

$$\int_0^R h(\chi^2) d\chi^2 \geq \sum_1^k w_i \int_0^{\frac{t_i^2}{v_i} \chi_i^2} h(\chi^2) d\chi^2 + \sum_1^l w'_j \int_0^{d^2} h(\chi^2) d\chi^2. \quad \dots (3.8)$$

It can be shown that
$$\int_0^\infty f(\chi_i^2) \left\{ \int_0^{\frac{t_i^2}{v_i} \chi_i^2} h(\chi^2) d\chi^2 \right\} d\chi_i^2 = \alpha \quad (i = 1, 2, \dots, k)$$

and
$$\int_0^{d^2} h(\chi^2) d\chi^2 = \alpha. \quad \dots (3.9)$$

As $\sum_1^k w_i + \sum_1^l w'_j = 1$, from (3.7), (3.8) and (3.9)

$$\text{prob} [(Y-M)^2 \leq \sum_1^k t_i^2 \lambda_i s_i^2 + d^2 \sum_1^l \theta_j \sigma_j^2] \geq \alpha. \quad \dots (3.10)$$

Result 1: Let three samples of N_1, N_2 and N_3 units be drawn from three normal populations with means m_1, m_2 and m_3 and variances not necessarily equal. Let $(\bar{x}_i, s_i^2)$ ($i = 1, 2, 3$) be respectively estimates for the population means and variances of the three populations. Suppose, based on previous experiments, we have estimates (i) $\bar{x}_{11}, s_{11}^2$; (ii) $\bar{x}_{21}$ and (iii) s_{31}^2 respectively for (i) mean and variance of the first population, (ii) mean of the second population and (iii) variance of the third population based on n_1, n_2 and n_3 units. Then a confidence interval for any linear function $\sum_1^3 c_i m_i$ of population means, with confidence coefficient not less than some pre-assigned probability α , may be built up from the relation:

$$\text{prob} \left[\left\{ \sum_1^3 c_i (\bar{x}_{i0} - m_i) \right\}^2 \leq \sum_1^3 \frac{t_i^2 c_i^2 s_{i0}^2}{N_{i0}} \right] \geq \alpha \quad \dots (3.11)$$

where

$$N_{10} = N_1 + n_1; \quad \bar{x}_{10} = \frac{N_1 \bar{x}_1 + n_1 \bar{x}_{11}}{N_{10}}$$

$$N_{20} = N_2 + n_2; \quad \bar{x}_{20} = \frac{N_2 \bar{x}_2 + n_2 \bar{x}_2}{N_{20}}$$

$$N_{30} = N_3; \quad \bar{x}_{30} = \bar{x}_3$$

$$s_{10}^2 = \frac{(N_1 - 1)s_1^2 + (n_1 - 1)s_{11}^2}{N_1 + n_1 - 2}$$

$$s_{20}^2 = s_2^2$$

$$s_{30}^2 = \frac{(N_3 - 1)s_3^2 + (n_3 - 1)s_{31}^2}{N_3 + n_3 - 2}$$

CONFIDENCE INTERVAL FOR TWO-MEANS PROBLEM

and t_i ($i = 1, 2, 3$) are tabulated values of Student's t -table of ν_i d.f. ($\nu_1 = N_1 + n_2 - 2$, $\nu_2 = N_2 - 1$ and $\nu_3 = N_3 + n_3 - 2$) corresponding to confidence coefficient α .

Relation (3.11) follows directly from Theorem 2.

Result 2: From two samples $(x_{ij}, y_{ij}; i = 1, 2; j = 1, 2, \dots, n_i)$ of n_i ($i = 1, 2$) units from two bivariate normal populations regression equations of y on x are estimated as:

$$\text{Population I : } Y_1 = \bar{y}_1 + b_1(X_1 - \bar{x}_1)$$

$$\text{Population II : } Y_2 = \bar{y}_2 + b_2(X_2 - \bar{x}_2).$$

If β_1 and β_2 denote respectively population values of regression coefficients of the two populations a confidence interval for $\beta_1 - \beta_2$ may be built up from the result :

$$\text{prob} \left[\{b_1 - b_2 - (\beta_1 - \beta_2)\}^2 \leq \frac{t_1^2 s_1^2}{S_1^2} + \frac{t_2^2 s_2^2}{S_2^2} \right] \geq \alpha \quad \dots (3.12)$$

where

$$S_i^2 = \sum_1^{n_i} (x_{ij} - \bar{x}_i)^2, \quad (i = 1, 2)$$

$$s_i^2 = \frac{\sum_1^{n_i} (y_{ij} - Y_{ij})^2}{n_i - 2}; \quad (i = 1, 2).$$

$$Y_{ij} = \bar{y}_i + b_i(x_{ij} - \bar{x}_i); \quad (i = 1, 2)$$

and t_i ($i = 1, 2$) are tabulated values of Student's t -table of $n_i - 2$ ($i = 1, 2$) d.f. corresponding to confidence coefficient α .

As b_1 and b_2 respectively are distributed normally about means β_1 and β_2 with variances say $\frac{\sigma_1^2}{S_1^2}$ and $\frac{\sigma_2^2}{S_2^2}$ and s_1^2 and s_2^2 are estimates of σ_1^2 and σ_2^2 with $n_1 - 2$ and $n_2 - 2$ d.f., the result follows from Theorem 2.

4. SOME RESULTS ON THE TWO-MEANS PROBLEM

In this section four results on the two-means problem would be presented as under :

(1) proof that restricting to class of functions of the form $A_1 s_1^2 + A_2 s_2^2$ the only function which with minimum values of A_1 and A_2 would satisfy the relation

$$\left\{ \sum_1^2 c_i (\bar{x}_i - m_i) \right\}^2 \leq A_1 s_1^2 + A_2 s_2^2$$

with probability not less than any pre-assigned probability α is

$$\frac{t_1^2 c_1^2 s_1^2}{n_1} + \frac{t_2^2 c_2^2 s_2^2}{n_2}$$

(2) numerical values of upper bounds of probability of the inequality

$$\left\{ \sum_1^2 c_i (\bar{x}_i - m_i) \right\}^2 \leq \frac{t_1^2 c_1^2 s_1^2}{n_1} + \frac{t_2^2 c_2^2 s_2^2}{n_2}$$

(3) proof that the confidence interval

$$\sum_1^2 c_i \bar{x}_i \pm \sqrt{\sum_1^2 t_i^2 c_i^2 s_i^2 / n_i} \text{ is unbiased}$$

(4) non-central confidence intervals.

Throughout this section and the following sections, unless otherwise stated, $\bar{x}_1$ and s_1^2 denote sample estimates of population mean m_1 and variance σ_1^2 based on a sample of size n_1 drawn from a normal population $N(m_1, \sigma_1^2)$. Also $\bar{x}_2$ and s_2^2 denote sample estimates of population mean m_2 and variance σ_2^2 based on a sample of size n_2 drawn from a normal population $N(m_2, \sigma_2^2)$. Also $\nu_1 (= n_1 - 1)$ and $\nu_2 (= n_2 - 1)$ denote degrees of freedom of s_1^2 and s_2^2 . Also c_1 and c_2 denote known constants, positive or negative, and α a pre-assigned probability level (between 0 and 1) and $t_i (i = 1, 2)$ are tabulated values of Student's t -table of $\nu_i (i = 1, 2)$ d.f. corresponding to confidence coefficient α .

Result 1: Let $P\{A_1, A_2\}$ denote the probability of the inequality

$$\left\{ \sum_1^2 c_i (\bar{x}_i - m_i) \right\}^2 \leq A_1 s_1^2 + A_2 s_2^2. \quad \dots (4.1)$$

It can be easily shown that $P\{A_1, A_2\}$ is equal to

$$\frac{1}{\Gamma(p_1)} \frac{1}{\Gamma(p_2)} \frac{1}{\Gamma(\frac{1}{2})} \int_0^\infty \int_0^\infty e^{-y_1 - y_2} \frac{y_1^{p_1-1}}{(y_1)} \frac{y_2^{p_2-1}}{(y_2)} \left\{ \int e^{-z} z^{-\frac{1}{2}} dz \right\} dy_1 dy_2 \quad \dots (4.2)$$

where

$$p_i = \nu_i/2; a_i = \frac{A_i/\nu_i}{c_i^2/n_i}; (i = 1, 2)$$

and

$$w_1 = \frac{c_1^2 \sigma_1^2 / n_1}{c_1^2 \sigma_1^2 / n_1 + c_2^2 \sigma_2^2 / n_2}.$$

Further, it can be shown that $P\{A_1, A_2\}$ as defined in (4.2) is continuous in w_1 for $0 \leq w_1 \leq 1$.

Let numerical values of A_1 and A_2 be so taken that

$$\left. \begin{aligned} A_1 &< \frac{t_1^2 c_1^2}{n_1} \\ A_2 &\geq \frac{t_2^2 c_2^2}{n_2} \end{aligned} \right\} \quad \dots (4.3)$$

and

CONFIDENCE INTERVAL FOR TWO-MEANS PROBLEM

With numerical values of A_1 and A_2 so determined by (4.3) as w_1 tends to unity $P\{A_1, A_2\}$ tends to α^1 where α^1 is defined as

$$\frac{1}{\Gamma(p_1)} \cdot \frac{1}{\Gamma(\frac{1}{2})} \cdot \int_0^\infty e^{-y_1} (y_1)^{p_1-1} \left\{ \int_0^{\frac{A_1/p_1}{c_1^2/n_1} y_1} e^{-z} z^{-\frac{1}{2}} dz \right\} dy_1 \quad \dots (4.4)$$

which is less than α .

Since $P\{A_1, A_2\}$ is continuous in w_1 it is possible to choose a value of w_1 near $w_1 = 1$ such that $P\{A_1, A_2\} < \alpha$. This means that even if A_2 is made arbitrarily large,

$$\text{prob} \left[\left\{ \sum_1^2 c_i(\bar{x}_i - m_i) \right\}^2 \leq A_1 s_1^2 + A_2 s_2^2 \right]$$

where $A_1 < \frac{t_1^2 c_1^2}{n_1}$ depending upon the value of w_1 , would be less than the pre-assigned probability level for some of the populations. Hence the only function of the form $A_1 s_1^2 + A_2 s_2^2$ which, with minimum values of A_1 and A_2 , would satisfy the relation

$$\left\{ \sum_1^2 c_i(\bar{x}_i - m_i) \right\}^2 \leq A_1 s_1^2 + A_2 s_2^2$$

with probability not less than α is

$$\frac{t_1^2 c_1^2 s_1^2}{n_1} + \frac{t_2^2 c_2^2 s_2^2}{n_2}$$

Result 2: From (4.2), $P \left\{ \frac{t_1^2 c_1^2}{n_1}, \frac{t_2^2 c_2^2}{n_2} \right\}$ standing for the probability of the

inequality

$$\left\{ \sum_1^2 c_i(\bar{x}_i - m_i) \right\}^2 \leq \frac{t_1^2 c_1^2 s_1^2}{n_1} + \frac{t_2^2 c_2^2 s_2^2}{n_2}$$

is equal to

$$\frac{1}{\Gamma(p_1)} \cdot \frac{1}{\Gamma(p_2)} \cdot \frac{1}{\Gamma(\frac{1}{2})} \int_0^\infty \int_0^\infty e^{-y_1 - y_2} (y_1)^{p_1-1} (y_2)^{p_2-1} \left\{ \int_0^{a_1 w_1 y_1 + a_2 w_2 y_2} e^{-z} z^{-\frac{1}{2}} dz \right\} dy_1 dy_2 \quad \dots (4.5)$$

where

$$p_i = v_i/2; \quad a_i = t_i^2/v_i; \quad (i = 1, 2)$$

$$w_1 = \frac{c_1^2 \sigma_1^2/n_1}{c_1^2 \sigma_1^2/n_1 + c_2^2 \sigma_2^2/n_2} \text{ and } w_2 = 1 - w_1.$$

$P \left\{ \frac{t_1^2 c_1^2}{n_1}, \frac{t_2^2 c_2^2}{n_2} \right\}$ is essentially a function of w_1 and it tends to α when w_1 tends to zero or unity. For values of w_1 between zero and unity $P \left\{ \frac{t_1^2 c_1^2}{n_1}, \frac{t_2^2 c_2^2}{n_2} \right\}$ is numerically

greater than α . Differentiating $P\left\{\frac{t_1^2 c_1^2}{n_1}, \frac{t_2^2 c_2^2}{n_2}\right\}$ with respect to w_1 within the sign of integration (such differentiation is permissible for $v_1, v_2 \geq 2$ and for $v_1, v_2 = 1$ if $0 < \epsilon \leq w_1 \leq 1 - \epsilon$) it can be shown that

$$\frac{d}{dw_1} P\left\{\frac{t_1^2 c_1^2}{n_1}, \frac{t_2^2 c_2^2}{n_2}\right\} = I_1 - I_2 \quad \dots (4.6)$$

where for (i) $a_2 w_2 \geq a_1 w_1$,

$$I_1 = K \cdot \frac{a_1 p_1}{1 + a_1 w_1} \cdot c_2^{-\frac{1}{2}} \cdot F\left\{\frac{1}{2}, p_1 + 1, p_1 + p_2 + 1; \frac{c_2 - c_1}{c_2}\right\} \quad \dots (4.7)$$

$$I_2 = K \cdot \frac{a_2 p_2}{1 + a_2 w_2} \cdot c_2^{-\frac{1}{2}} \cdot F\left\{\frac{1}{2}, p_1, p_1 + p_2 + 1; \frac{c_2 - c_1}{c_2}\right\} \quad \dots (4.8)$$

and for (ii) $a_1 w_1 \geq a_2 w_2$,

$$I_1 = K \cdot \frac{a_1 p_1}{1 + a_1 w_1} \cdot c_1^{-\frac{1}{2}} \cdot F\left\{\frac{1}{2}, p_2, p_1 + p_2 + 1; \frac{c_1 - c_2}{c_1}\right\} \quad \dots (4.9)$$

$$I_2 = K \cdot \frac{a_2 p_2}{1 + a_2 w_2} \cdot c_1^{-\frac{1}{2}} \cdot F\left\{\frac{1}{2}, p_2 + 1, p_1 + p_2 + 1; \frac{c_1 - c_2}{c_1}\right\} \quad \dots (4.10)$$

where $c_1 = \frac{a_1 w_1}{1 + a_1 w_1}$; $c_2 = \frac{a_2 w_2}{1 + a_2 w_2}$; and K is a function of p_i, a_i and w_i ($i = 1, 2$).

If $v_1 = v_2 = v$, so that $a_1 = a_2 = a_0$, from (4.7) and (4.8) for the case $a_0 w_2 \geq a_0 w_1$ (i.e. $w_1 \leq \frac{1}{2}$)

$$I_1 = K' \cdot \frac{1}{1 + a_0 w_1} \cdot F\left\{\frac{1}{2}, \frac{v}{2} + 1, v + 1; \frac{w_2 - w_1}{w_2(1 + a_0 w_1)}\right\} \quad \dots (4.11)$$

$$I_2 = K' \cdot \frac{1}{1 + a_0 w_2} \cdot F\left\{\frac{1}{2}, \frac{v}{2}, v + 1; \frac{w_2 - w_1}{w_2(1 + a_0 w_1)}\right\} \quad \dots (4.12)$$

As

$$F\left\{\frac{1}{2}, \frac{v}{2} + 1, v + 1; z\right\} > F\left\{\frac{1}{2}, \frac{v}{2}, v + 1; z\right\}, \quad (z > 0)$$

from (4.11) and (4.12) it follows for $w_1 < \frac{1}{2}$,

$$I_1 > I_2 \quad \dots (4.13)$$

which means $P\left\{\frac{t_1^2 c_1^2}{n_1}, \frac{t_2^2 c_2^2}{n_2}\right\}$ increases as w_1 increases. Further from (4.11) and

(4.12) it is evident that at $w_1 = \frac{1}{2}$

$$I_1 = I_2 \quad \dots (4.14)$$

CONFIDENCE INTERVAL FOR TWO-MEANS PROBLEM

Also from (4.9) and (4.10) for the case $a_0 w_1 > a_0 w_2$ (i.e. $w_1 > \frac{1}{2}$) it can be shown that

I_2 is greater than I_1 . $P \left\{ \frac{t_0^2 c_1^2}{n}, \frac{t_0^2 c_2^2}{n} \right\}$ for variation in w_1 takes a maximum value at

$w_1 = \frac{1}{2}$. In Table 1 below maximum value of $P \left\{ \frac{t_0^2 c_1^2}{n}, \frac{t_0^2 c_2^2}{n} \right\}$ for suitable values of

$\nu (= n-1)$ and α (i) 0.90; (ii) 0.95 and (iii) 0.98 are presented. The values have been worked out from Table 9 (Probability Integral $P(t/\nu)$ of the t -distribution) of *Biometrika Tables*, Volume I, using the relation :

$$P \left\{ \frac{t_0^2 c_1^2}{n}, \frac{t_0^2 c_2^2}{n} \right\}_{w_1 = \frac{1}{2}} = \int_{-t_0}^{t_0} f(t/2\nu) dt$$

where t_0 satisfies the relation

$$\int_{-t_0}^{t_0} f(t/\nu) dt = \alpha.$$

Also, occasionally, Incomplete Beta Function tables were used.

TABLE 1. MAXIMUM VALUES OF $P \left\{ \frac{t_0^2 c_1^2}{n}, \frac{t_0^2 c_2^2}{n} \right\}$

FOR VARIATION IN w_1

$\nu = n-1$	α -value		
	0.90	0.95	0.98
(0)	(1)	(2)	(3)
1	.9758	.9939	.9990
2	.9567	.9873	.9978
3	.9430	.9809	.9961
4	.9342	.9759	.9943
5	.9283	.9721	.9928
6	.9239	.9691	.9915
8	.9183	.9651	.9895
10	.9149	.9623	.9880
12	.9125	.9605	.9869
14	.9108	.9591	.9860
16	.9090	.9579	.9854
18	.9080	.9572	.9848
20	.9077	.9565	.9844
22	.9062	.9559	.9840
24	.9055	.9554	.9836
30	.9051	.9542	.9830

Result 3: The confidence interval

$$\Sigma c_i \bar{x}_i \pm \sqrt{\frac{t_1^2 c_1^2 s_1^2}{n_1} + \frac{t_2^2 c_2^2 s_2^2}{n_2}} \quad \dots \quad (4.15)$$

is unbiased.

Let

$$\left. \begin{aligned} P_0 &\equiv \text{prob} \left[\left\{ \sum_1^2 c_i \bar{x}_i - \sum_1^2 c_i m_i \right\}^2 \leq \sum_1^2 \frac{t_i^2 c_i^2 s_i^2}{n_i} \right] \\ \text{and} \quad P_1 &\equiv \text{prob} \left[\left\{ \sum_1^2 c_i \bar{x}_i - M' \right\}^2 \leq \sum_1^2 \frac{t_i^2 c_i^2 s_i^2}{n_i} \right] \end{aligned} \right\} \dots (4.16)$$

Since for fixed (s_1, s_2) , $\sum_1^2 c_i \bar{x}_i$ is distributed normally about mean $\sum_1^2 c_i m_i$ with variance, say θ^2 ,

$$P_0 = \int_0^\infty \int_0^\infty g(s_1) g(s_2) \left\{ \int_{-T/\theta}^{T/\theta} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \right\} ds_1 ds_2$$

$$\text{and} \quad P_1 = \int_0^\infty \int_0^\infty g(s_1) g(s_2) \left\{ \int_{-T/\theta}^{T/\theta} \frac{1}{\sqrt{2\pi}} e^{-(u-M'')^2/2} du \right\} ds_1 ds_2$$

where

$g(s_1)$ = frequency function of s_1

$g(s_2)$ = „ „ of s_2

$$T = \sqrt{\sum_1^2 \frac{t_i^2 c_i^2 s_i^2}{n_i}} \text{ and } M'' = \sum_1^2 c_i m_i - M'.$$

As

$$\int_{-d}^d e^{-u^2/2} du > \int_{-d}^d e^{-(u-M'')^2/2} du,$$

unbiasedness follows.

Result 4: The confidence interval so far considered for the two-means problem is central by nature. Non-central confidence interval with confidence coefficient not less than any pre-assigned probability level is, however, possible. Consider the confidence interval :

$$\left\{ -T_1(s_1^2, s_2^2) \leq \sum_1^2 c_i (\bar{x}_i - m_i) \leq T_2(s_1^2, s_2^2) \right\} \dots (4.17)$$

where

$$T_1(s_1^2, s_2^2) \equiv \sqrt{\frac{t_{11}^2 c_1^2 s_1^2}{n_1} + \frac{t_{21}^2 c_2^2 s_2^2}{n_2}}$$

$$T_2(s_1^2, s_2^2) \equiv \sqrt{\frac{t_{12}^2 c_1^2 s_1^2}{n_1} + \frac{t_{22}^2 c_2^2 s_2^2}{n_2}}$$

CONFIDENCE INTERVAL FOR TWO-MEANS PROBLEM

where t_{ij} ($i = 1, 2, j = 1, 2$) have been so determined that

$$\left. \begin{aligned} \int_{-t_{11}}^0 f(t/\nu_1) dt &= \alpha_{11}; \quad \int_0^{t_{12}} f(t/\nu_1) dt = \alpha_{12}; \\ \int_{-t_{21}}^0 f(t/\nu_2) dt &= \alpha_{21}; \quad \int_0^{t_{22}} f(t/\nu_2) dt = \alpha_{22}; \end{aligned} \right\} \dots (4.18)$$

where

$$t_{ij} \geq 0 (i, j = 1, 2)$$

and

$$\alpha_{11} + \alpha_{12} = \alpha_{21} + \alpha_{22} = \alpha.$$

Now $\text{prob} \{-T_1 \leq \Sigma c_i(\bar{x}_i - m_i) \leq T_2\}$

$$\begin{aligned} &= \int_0^\infty \int_0^\infty g(s_1) g(s_2) \left\{ \int_{-T_1/\theta}^{T_2/\theta} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \right\} ds_1 ds_2 \\ &= \frac{1}{2} \int_0^\infty \int_0^\infty g(s_1) g(s_2) \left\{ \int_0^{T_1^2/\theta^2} h(\chi^2) d\chi^2 + \int_0^{T_2^2/\theta^2} h(\chi^2) d\chi^2 \right\} ds_1 ds_2 \dots (4.19) \end{aligned}$$

where

$g(s_1)$ = frequency function of s_1

$g(s_2)$ = " " of s_2

$h(\chi^2)$ = " " of χ^2 variate with 1.d.f.

and

$$\theta^2 = \frac{c_1^2 \sigma_1^2}{n_1} + \frac{c_2^2 \sigma_2^2}{n_2}.$$

$$\text{Now } \int_0^{T_1^2/\theta^2} h(\chi^2) d\chi^2 \geq w_1 \int_0^{t_1^2 s_1^2 / \sigma_1^2} h(\chi^2) d\chi^2 + (1-w_1) \int_0^{t_2^2 s_2^2 / \sigma_2^2} h(\chi^2) d\chi^2 \dots (4.20)$$

where

$$w_1 = \frac{c_1^2 \sigma_1^2 / n_1}{c_1^2 \sigma_1^2 / n_1 + c_2^2 \sigma_2^2 / n_2}, \quad (i = 1, 2).$$

From (4.19) and (4.20)

$$\text{prob} [-T_1 \leq \Sigma c_i(\bar{x}_i - m_i) \leq T_2] \geq \frac{1}{2} \{w_1(\alpha_{11} + \alpha_{12}) + (1-w_1)(\alpha_{21} + \alpha_{22})\} = \alpha.$$

5. TEST OF HYPOTHESIS

The problem of the two-means has so far been considered from the approach of estimation by confidence interval corresponding to the region R

$$\left\{ \sum_1^2 c_i(\bar{x}_i - m_i) \right\}^2 \leq \frac{t_1^2 c_1^2 s_1^2}{n_1} + \frac{t_2^2 c_2^2 s_2^2}{n_2}.$$

The region R' , complementary to the region of acceptance R , may be used as a critical region for testing hypothesis H regarding population means. For example,

if it is required to test the hypothesis H that $m_1 = m_2$ the corresponding statistic T may be defined as

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{t_1^2 s_1^2}{n_1} + \frac{t_2^2 s_2^2}{n_2}}},$$

t_1 and t_2 being suitably chosen. If such tests are applied, the first kind of error, i.e., the probability of rejection of the hypothesis when true, will not be exactly equal to an assigned value $1-\alpha$ but would depend upon w_1 or the variances of the two populations. For w_1 tending to zero or unity, the error of the first kind would tend to the assigned value $1-\alpha$. Hence, one can go in for such tests, if one is prepared to accept tests whose first kind of error would not be exactly equal to a given value $1-\alpha$ but would depend upon the values of the free parameters (here σ_1^2, σ_2^2) in such a way that the first kind of error would always be less than or equal to $1-\alpha$. Such tests, however, would be unbiased as the complementary region R , the region of acceptance, as proved earlier is unbiased

Two examples have been considered below. Example 1 has been taken from *Biometrika Tables*, Volume I, 1954 and Example 2 from *Statistical Tables* by Fisher and Yates.

Example 1: Two samples of sizes $n_1 = 10$ and $n_2 = 15$ furnish the following estimates :

population	mean	variance
I	73.4	51
II	47.1	141

To test the hypothesis about the equality of population means with maximum value of error of the first kind fixed at (i) 0.10, (ii) 0.05 and (iii) 0.02 three statistics (i) T_1 , (ii) T_2 and (iii) T_3 respectively may be computed as under :

$$T_j = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{t_{1j}^2 s_1^2}{n_1} + \frac{t_{2j}^2 s_2^2}{n_2}}} \quad (j = 1, 2, 3); \quad T_2 = \frac{73.4 - 47.1}{\sqrt{69.34}} = \frac{26.3}{8.33} = 3.16;$$

$$T_1 = \frac{73.4 - 47.1}{\sqrt{46.29}} = \frac{26.3}{6.80} = 3.87; \quad T_3 = \frac{73.4 - 47.1}{\sqrt{105.31}} = \frac{26.3}{10.26} = 2.56;$$

where t_{ij} are $100.\alpha_j$ percentile points of Student's t -table of v_i d.f. with $v_1 = 9, v_2 = 14$ and $\alpha_1 = 0.10, \alpha_2 = 0.05$ and $\alpha_3 = 0.02$. All the statistics are numerically greater than unity. It is seen that with maximum value of the error of the first kind fixed even at 0.02 the hypothesis cannot be accepted.

CONFIDENCE INTERVAL FOR TWO-MEANS PROBLEM

Apart from the question of testing of hypothesis, confidence intervals I_1 , I_2 and I_3 for the difference of population means with confidence coefficient respectively not less than (i) 0.90, (ii) 0.95 and (iii) 0.98 may be built up as

$$(i) \quad I_1 \rightarrow 26.3 \pm 6.80 \rightarrow (19.50 \rightarrow 33.10)$$

$$(ii) \quad I_2 \rightarrow 26.3 \pm 8.33 \rightarrow (17.97 \rightarrow 34.63)$$

$$(iii) \quad I_3 \rightarrow 26.3 \pm 10.26 \rightarrow (16.04 \rightarrow 36.56)$$

90 per cent confidence limits, according to Welch's solution for this case is 19.8 to 32.8 which corresponds to interval I_1 having a range 19.5 to 33.1.

Example 2 : A physical constant evaluated by a new method gives a mean of twelve determinations,

$$\bar{x}_1 = 4.77383$$

and that the sum of squares of the deviations of these values from their mean is

$$\sum_j (\bar{x}_{1j} - \bar{x}_1)^2 = 0.11580 \times 10^{-1}$$

so that from 11 degrees of freedom the variance of the mean is estimated to be

$$s_1^2(\bar{x}_1) = 0.8773 \times 10^{-4}$$

and the estimated standard deviation of the mean is

$$s_1(\bar{x}_1) = .9366 \times 10^{-2}.$$

Numerous previous determinations, using different methods, have given the value,

$$\bar{x}_2 = 4.744$$

where $\bar{x}_2$ has a standard error based on a large number of degrees of freedom with $s_2(\bar{x}_2) = .00382$.

To test the hypothesis about the equality of population means with maximum value of error of the first kind fixed at (i) 0.10, (ii) 0.05 and (iii) 0.02 three statistics (i) T_1 , (ii) T_2 and (iii) T_3 respectively may be computed as under :

$$T_j = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{t_{1j}^2 s_1^2(\bar{x}_1) + t_{2j}^2 s_2^2(\bar{x}_2)}}; \quad T_2 = \frac{.02983}{10^{-2} \times \sqrt{4.8105}} = \frac{2.983}{2.19} = 1.36;$$

$$T_1 = \frac{.02983}{10^{-2} \times \sqrt{3.2246}} = \frac{2.983}{1.80} = 1.66; \quad T_3 = \frac{.02893}{10^{-2} \times \sqrt{7.2704}} = \frac{2.983}{2.70} = 1.11;$$

where t_{ij} are $100 \cdot \alpha_j$ percentile points of Student's t -table of ν_i d.f. with $\nu_1 = 11$ and $\nu_2 = \infty$ and $\alpha_1 = 0.10$, $\alpha_2 = 0.05$ and $\alpha_3 = 0.02$. All the statistics are numerically greater than unity. It is seen that with maximum value of the error of the first kind fixed even at 0.02 the hypothesis cannot be accepted.

6. BEHRENS-FISHER AND WELCH'S SOLUTION

Fisher (1935) indicated that by means of fiducial argument in statistical inference given two samples of $n_i (i = 1, 2)$ units from two normal populations with $\bar{x}_i$ and $s_i^2 (i = 1, 2)$ as sample estimates of population means and variances, the relation

$$d = \frac{(\bar{x}_1 - m_1) - (\bar{x}_2 - m_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = u_1 \sin \theta - u_2 \cos \theta \quad \dots (6.1)$$

where $u_i = \frac{\bar{x}_i - m_i}{s_i/\sqrt{n_i}} (i = 1, 2)$ and $\tan \theta = \frac{s_1/\sqrt{n_1}}{s_2/\sqrt{n_2}}$

originally due to Behrens (1929), could be used to test the hypothesis that $d (= m_1 - m_2)$ has the value zero. Sukhatme in 1938 published critical values of Behrens-Fisher test as defined in (6.1) for 5 per cent level of significance for $\nu_1, \nu_2 = 6, 8, 12, 24$ and ∞ for $\theta = 0^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ, 90^\circ$ where $\tan \theta$ is $\frac{s_1/\sqrt{n_1}}{s_2/\sqrt{n_2}}$. Further critical values for 1 per cent level of significance for $\nu_1, \nu_2 = 6, 8, 12, 24$ and ∞ were published later. To calculate critical values of (6.1) Sukhatme assumed that for given value of θ, u_1 and u_2 were independently distributed as Student's t -variate with $n_i - 1 (i = 1, 2)$ d.f. Critical values of Behrens-Fisher test for small odd degrees of freedom $\nu_1, \nu_2 = 1, 3, 5$ and 7 were published by Fisher and Healy in 1956.

Critical values of Behrens-Fisher test have been tabulated for different values of θ where $\tan \theta = \frac{s_1/\sqrt{n_1}}{s_2/\sqrt{n_2}}$. For $\theta = 0^\circ$, critical values of Behrens-Fisher test is equal to t -values of Student's t -table with ν_2 d.f. Also for $\theta = 90^\circ$, critical values of Behrens-Fisher test is exactly equal to t -values of Student's t -table with ν_1 d.f. For intermediate values of θ , critical values of the Behrens-Fisher test for $\nu_1 = \nu_2 = \nu$ and $\nu = 6, 8, 12$, and 24 is numerically less than tabulated critical values for $\theta = 0^\circ$ (or 90° , which are numerically equal in such cases) for 5 per cent and 1 per cent level of significance. For $\nu_1 \neq \nu_2$ and $\nu_1, \nu_2 = 6, 8, 12, 24$ and ∞ critical values of Behrens-Fisher test for intermediate values of θ usually lies in between tabulated critical values for $\theta = 0^\circ$ and $\theta = 90^\circ$ and is occasionally less than both of them. For $\nu_1 = \nu_2 = 1$ critical values of the test for intermediate values θ are, however, numerically higher than corresponding critical values for $\theta = 0^\circ$ (or 90°) for 10 per cent, 5 per cent, 2 per cent and 1 per cent level of significance.

Welch (1947) considered the problem of finding a function which is such that

$$\text{prob} \left[\frac{|y - \eta|}{\sqrt{\sum_{i=1}^k \lambda_i s_i^2}} \leq V(s_1^2, s_2^2, \dots, s_k^2) \right] = \alpha \quad \dots (6.2)$$

CONFIDENCE INTERVAL FOR TWO-MEANS PROBLEM

for all values of $\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2$, where y is a normal variate distributed about mean η with variance $\sum_1^k \lambda_i \sigma_i^2$ and s_i^2 are independent estimates of σ_i^2 and λ_i are known positive constants ($i = 1, 2, \dots, k$). For $y \equiv \bar{x}_1 - \bar{x}_2$ Welch's problem is to find $V(s_1^2, s_2^2)$ which is such that

$$\text{prob} \left[\frac{|\bar{x}_1 - \bar{x}_2 - (m_1 - m_2)|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leq V(s_1^2, s_2^2) \right] = \alpha. \quad \dots (6.3)$$

Wilks (1940) stated that an exact solution of the form (6.3) is not possible; no proof has, however, been published. Welch has put forward a series for the case (6.2) (which includes the case (6.3) as well) as under:

$$\begin{aligned} V(s_1^2, s_2^2, \dots, s_k^2) = & \xi + \frac{\xi(1+\xi^2)}{4} \left(\sum_1^k c_i^2/f_i \right) - \frac{\xi(1+\xi^2)}{2} \left(\sum_1^k c_i^2/f_i^2 \right) \\ & + \frac{\xi(3+5\xi^2+\xi^4)}{3} \left(\sum_1^k c_i^3/f_i^2 \right) \text{ etc.} \quad \dots (6.4) \end{aligned}$$

where

$$c_i = \lambda_i s_i^2 / \sum_1^k \lambda_i s_i^2 ;$$

$$f_i = \text{d.f. of } s_i^2 ; (i = 1, 2, \dots, k)$$

and ξ is tabulated value of normal probability table so that

$$\frac{1}{\sqrt{2\pi}} \int_{-\xi}^{\xi} e^{-\frac{y^2}{2}} dy = \alpha.$$

In the words of Bartlett (1956) "there is a permissible criticism of Welch's solution, namely, that the existence of an exact solution in his sense has never been rigorously established." According to Wallace (1958) "it is still not known whether a non-randomized similar level α test exists."

Critical values of Welch's solution for the two sample case only have been calculated by Aspin and are given in *Biometrika Tables*, Volume I (Table No. 11, 1954 print). The values given in *Biometrika Tables* cover the range (i) $v_1, v_2 = 6, 8, 10, 15, 20$ and ∞ for $\alpha = 0.90$ and (ii) $v_1, v_2 = 10, 12, 15, 20, 30$ and ∞ for $\alpha = 0.98$. Also further critical values for $\alpha = 0.95$ and 0.99 are given in Welch, Trickett and James (1956). Critical values of Welch's solution for very small degrees of freedom ($v_1, v_2 \leq 5$) have not been published.

Critical values $V(c, v_1, v_2, \alpha)$ of Welch's solution have been tabulated for given v_1, v_2 and α for the ratio $c = \frac{\lambda_1 s_1^2}{\lambda_1 s_1^2 + \lambda_2 s_2^2}$. This, of course, does not mean that critical values of Welch's solution refer to sub-sets having observed variance ratios. Both Welch's solution and the present solution refer to unrestricted variation of $\bar{x}_1$,

$\bar{x}_2$, s_1^2 and s_2^2 . For $c = 0$ critical values of Welch's solution are exactly equal to t -values of the Students t -values with ν_2 degrees of freedom. Also for $c = 1$ critical values of Welch's solution are exactly equal to t -values of the Student's t -values with ν_1 degrees of freedom. For intermediate values of c , $V(c, \nu_1, \nu_2, \alpha)$ numerically lies in between $V(0, \nu_1, \nu_2, \alpha)$ and $V(1, \nu_1, \nu_2, \alpha)$.

To compare the solution derived in this paper with Welch's solution, the solution derived in this paper can be written as under :

$$\text{prob} \left[\frac{(y-\eta)^2}{\lambda_1 s_1^2 + \lambda_2 s_2^2} \leq \frac{t_1^2 \lambda_1 s_1^2 + t_2^2 \lambda_2 s_2^2}{\lambda_1 s_1^2 + \lambda_2 s_2^2} \right] \geq \alpha. \quad \dots (6.5)$$

Using relation (6.5), critical values of (i) Welch's solution and (ii) the present solution for the cases (i) $\alpha = 0.95$; $\nu_2 = 8$; $\nu_1 = 8, 12$ and ∞ and (ii) $\alpha = 0.99$; $\nu_2 = 12$; $\nu_1 = 12$ and ∞ have been given in Tables 2 and 3. Critical values of the present solution as given in column (4) of Tables 2 and 3 have been worked out from the relation:

$$\sqrt{t_1^2 c + t_2^2 (1-c)}$$

where

$$c = \frac{\lambda_1 s_1^2}{\lambda_1 s_1^2 + \lambda_2 s_2^2}$$

for different values of c as indicated in column (1) of Table 2.

It is seen from Tables 2 and 3 that for $c = 0$ or 1, there is no difference between the critical values of the two solutions. For intermediate values of c there are differences in the values but the magnitude of the differences falls off as the number of degrees of freedom increases. Maximum magnitude of the differences expressed as percentage critical values of the Welch's solution are of the order of (i) 11 per cent (ii) 7 per cent and (iii) 4 per cent respectively for the cases $\nu_1 =$ (i) 8, (ii) 12 and (iii) ∞ with $\nu_2 = 8$ and $\alpha = 0.95$. Difference in the mathematical expectation of the lengths of the two confidence intervals, which would be a weighted average of the difference in the lengths of the two intervals at different c values, properly weighted by the frequency of occurrence of c , would, however, be smaller than the maximum magnitude of the difference.

A comparison of the critical values of the present solution with the critical values of the Behrens-test as given in *Statistical Tables* (by Fisher and Yates) may be of interest. This comparison, however, would not be strictly valid because the present solution is based on the theory of confidence interval whereas the tabulated values as given in statistical tables are based on the theory of fiducial inference. Further, presumably the Behrens-Fisher test calculates critical values restricting to sub-sets

CONFIDENCE INTERVAL FOR TWO-MEANS PROBLEM

TABLE 2. CRITICAL VALUES OF (i) WELCH (ii) BEHRENS-FISHER AND
(iii) PRESENT SOLUTION FOR $\nu_2 = 8$; $\nu_1 = 8, 12$ AND ∞
FOR 5 PER CENT LEVEL OF SIGNIFICANCE

c	Welch's solution	Behrens-Fisher test*	present solution	difference col. (4)-col. (2)	difference col. (4)-col. (3)
(1)	(2)	(3)	(4)	(5)	(6)
$\nu_2 = 8; \nu_1 = 8$					
0.0	2.31	2.31	2.31	0.00	0.00
0.1	2.25	2.30	2.31	0.06	0.01
0.2	2.20	2.30	2.31	0.11	0.01
0.3	2.14	2.29	2.31	0.17	0.02
0.4	2.10	2.29	2.31	0.21	0.02
0.5	2.08	2.29	2.31	0.23	0.02
0.6	2.10	2.29	2.31	0.21	0.02
0.7	2.14	2.29	2.31	0.17	0.02
0.8	2.20	2.30	2.31	0.11	0.01
0.9	2.25	2.30	2.31	0.06	0.01
1.0	2.31	2.31	2.31	0.00	0.00
$\nu_2 = 8; \nu_1 = 12$					
0.0	2.31	2.31	2.31	0.00	0.00
0.1	2.25	2.29	2.29	0.04	0.00
0.2	2.20	2.27	2.28	0.08	0.01
0.3	2.15	2.26	2.27	0.12	0.01
0.4	2.10	2.24	2.26	0.16	0.02
0.5	2.07	2.23	2.24	0.17	0.01
0.6	2.07	2.22	2.23	0.16	0.01
0.7	2.08	2.21	2.22	0.14	0.01
0.8	2.11	2.20	2.20	0.09	0.00
0.9	2.14	2.19	2.19	0.05	0.00
1.0	2.18	2.18	2.18	0.00	0.00
$\nu_2 = 8; \nu_1 = \infty$					
0.0	2.31	2.31	2.31	0.00	0.00
0.1	2.25	2.27	2.27	0.02	0.00
0.2	2.20	2.23	2.24	0.04	0.01
0.3	2.14	2.20	2.21	0.07	0.01
0.4	2.09	2.16	2.17	0.08	0.01
0.5	2.05	2.13	2.14	0.09	0.01
0.6	2.01	2.09	2.10	0.09	0.01
0.7	1.99	2.06	2.07	0.08	0.01
0.8	1.97	2.03	2.03	0.06	0.00
0.9	1.96	1.99	2.00	0.04	0.01
1.0	1.96	1.96	1.96	0.00	0.00

* Values have been worked out from tabulated values as given in Table VI, *Statistical Tables* by Fisher and Yates, by interpolation.

TABLE 3. CRITICAL VALUES OF (i) WELCH (ii) BEHRENS-FISHER AND
(iii) PRESENT SOLUTION FOR $\nu_2 = 12$; $\nu_1 = 12$ AND ∞
FOR 1 PER CENT LEVEL OF SIGNIFICANCE

c	Welch's solution	Behrens-Fisher test*	present solution	difference col. (4)-col. (2)	difference col. (4)-col. (3)
(1)	(2)	(3)	(4)	(5)	(6)
$\nu_2 = 12$; $\nu_1 = 12$					
0.0	3.05	3.06	3.06	0.01	0.00
0.1	2.98	3.02	3.06	0.08	0.04
0.2	2.91	2.99	3.06	0.15	0.07
0.3	2.84	2.97	3.06	0.22	0.09
0.4	2.78	2.96	3.06	0.28	0.10
0.5	2.76	2.95	3.06	0.30	0.11
0.6	2.78	2.96	3.06	0.28	0.10
0.7	2.84	2.97	3.06	0.22	0.09
0.8	2.91	2.99	3.06	0.15	0.07
0.9	2.98	3.02	3.06	0.08	0.04
1.0	3.05	3.06	3.06	0.01	0.00
$\nu_2 = 12$; $\nu_1 = \infty$					
0.0	3.05	3.06	3.06	0.01	0.00
0.1	2.98	3.00	3.01	0.03	0.01
0.2	2.91	2.94	2.96	0.05	0.02
0.3	2.84	2.88	2.92	0.08	0.04
0.4	2.77	2.83	2.87	0.10	0.04
0.5	2.71	2.78	2.83	0.12	0.05
0.6	2.65	2.73	2.78	0.13	0.05
0.7	2.62	2.68	2.73	0.11	0.05
0.8	2.59	2.64	2.68	0.09	0.04
0.9	2.58	2.61	2.63	0.05	0.02
1.0	2.58	2.58	2.58	0.00	0.00

* Values have been worked out from tabulated values as given in Table VI, *Statistical Tables* by Fisher and Yates, by interpolation.

having observed values $\frac{s_1/\sqrt{n_1}}{s_2/\sqrt{n_2}}$ whereas in the present case critical values refer to unrestricted variation of the four sample estimates $\bar{x}_1$, $\bar{x}_2$, s_1^2 and s_2^2 .

Bearing in mind the broad limitations of comparing the critical values of the present solution with the critical values of Behrens-Fisher test, in column (6) of Tables 2 and 3 differences in critical values of the two solutions have been shown. It is seen that the differences in the critical values are small.

Further comparison of critical values of the present solution with the critical values of the Behrens-Fisher test for small d.f. ν_1 , $\nu_2 = 1, 3$ and 5 has been done in

CONFIDENCE INTERVAL FOR TWO-MEANS PROBLEM

Table 4. [Welch's solution has not been considered because critical values of Welch's solution for $\nu_1, \nu_2 \leq 5$ have not been published.] It is seen from columns (4) and (7) that excepting the cases (i) $\nu_1 = \nu_2 = 1$; (ii) $\nu_1 = 1; \nu_2 = 3$ differences in the critical values are usually small.

TABLE 4. CRITICAL VALUES OF (i) BEHRENS-FISHER TEST AND
(ii) PRESENT SOLUTION FOR SMALL ODD DEGREES OF
FREEDOM FOR 5 PER CENT LEVEL OF SIGNIFICANCE

θ	Behrens- Fisher test*	present solution	difference col.(3)-col. (2)	Behrens- Fisher test ¹	present solution	difference col.(6)-col. (5)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\nu_1 = 1; \nu_2 = 1$				$\nu_1 = 1; \nu_2 = 3$		
0°	12.706	12.706	0.000	3.182	3.182	.000
15°	15.562	12.706	-2.856	4.960	4.501	-.459
30°	17.357	12.706	-4.651	7.123	6.925	-.198
45°	17.969	12.706	-5.263	9.303	9.262	-.041
60°	17.357	12.706	-4.651	11.112	11.118	.006
75°	15.562	12.706	-2.856	12.294	12.300	.006
90°	12.706	12.706	0.000	12.706	12.706	.000
$\nu_1 = 1; \nu_2 = 5$				$\nu_1 = 3; \nu_2 = 3$		
0°	2.571	2.571	.000	3.182	3.182	.000
15°	4.218	4.121	-.097	3.191	3.182	-.009
30°	6.636	6.732	.096	3.225	3.182	-.043
45°	9.090	9.166	.076	3.244	3.182	-.062
60°	11.043	11.077	.034	3.225	3.182	-.043
75°	12.282	12.292	.010	3.191	3.182	-.009
90°	12.706	12.706	.000	3.182	3.182	.000
$\nu_1 = 3; \nu_2 = 5$				$\nu_1 = 5; \nu_2 = 5$		
0°	2.571	.571	.000	2.571	2.571	.000
15°	2.626	2.617	-.009	2.564	2.571	.007
30°	2.756	2.737	-.019	2.562	2.571	.009
45°	2.897	2.893	-.004	2.565	2.571	.006
60°	3.026	3.041	.015	2.562	2.572	.009
75°	3.134	3.145	.011	2.564	2.571	.007
90°	3.182	3.182	.000	2.571	2.571	.000

* Values taken from Table VI, *Statistical Tables* (1957 edition) by Fisher and Yates.

7. CONCLUSION

Given two samples of n_i ($i = 1, 2$) units from two normal populations, having variances not necessarily equal, a confidence interval for any linear function of population means in terms of sample estimates of population means and variances and tabulated values of Student's t -tables is possible. If the population variances are unknown the only function of the form $A_1 s_1^2 + A_2 s_2^2$ which with minimum values of A_1 and A_2 would satisfy the relation

$$\left\{ \sum_1^2 c_i (\bar{x}_i - m_i) \right\}^2 \leq A_1 s_1^2 + A_2 s_2^2$$

with probability not less than α is $\frac{t_1^2 c_1^2 s_1^2}{n_1} + \frac{t_2^2 c_2^2 s_2^2}{n_2}$. Further with maximum value of the error of the first kind (probability of rejection of hypothesis when true) fixed at any given value any hypothesis regarding the equality of population means (or any linear function of population means) of the two populations can be tested. Such tests are unbiased.

ACKNOWLEDGEMENT

Acknowledgements are due to Professor P. C. Mahalanobis, and to Dr. Nripen Sen of the Department of Applied Mathematics, Calcutta University, and Dr. Sanat Kumar Bose of Presidency College, Calcutta.

REFERENCES

- ASPIN, A. A. (1949): Tables for use in comparisons whose accuracy involves two variances separately estimated (with an Appendix by B. I. Welch). *Biometrika*, **36**, 290-296.
- BANERJEE, S. K. (1960): Approximate confidence interval for linear functions of means of k -populations when the population variances are not equal. *Sankhyā*, **22**, 357-358.
- BEHRENS, W. V. (1928): *Landw. Jb.*, **68**, 807-837.
- BARTLETT, M. S. (1956): Comment on Sir Roland Fisher's paper "On a test of significance in *Biometrika Tables* (No. 11)", *J. Roy. Stat. Soc.*, **B**, **18**, 295-296.
- FISHER, R. A. (1935): The fiducial argument in statistical inference. *Ann. Eugen.*, **6**, 391-398.
- FISHER, R. A. and HEALY, M. J. R. (1956): New tables of Behrens test of significance. *J. Roy. Stat. Soc.*, **B**, **18**, 212-216.
- FISHER, R. A. and YATES, F. (1957): *Statistical Tables*, Oliver and Boyd.
- PEARSON, E. S. and HARTLEY, H. O. (eds): *Biometrika Tables for Statisticians*, 1, 1954 print, Cambridge University Press.
- SUKHATME, P. V. (1938): On Fisher and Behrens test of significance for the difference in means of two normal samples. *Sankhyā*, **4**, 39-48.
- WALLACE, D. L. (1958): Asymptotic approximations. *Ann. Math. Stat.*, **29**, 635-654.
- WELCH, B. L. (1947): The significance of the difference between two means when the population variances are unequal. *Biometrika*, **29**, 350-362.
- WELCH, B. L. and TRICKETT, W. H. (1954): On the comparison of two means: further discussion of iterative methods for calculation of tables. *Biometrika*, **41**, 361-374.
- WELCH, B. L., TRICKETT, W. H. and JAMES G. S. (1956): Further critical values of the two-means problem. *Biometrika*, **43**, 203-205.
- WILKS, S. S. (1940): On the problem of two samples from normal populations with unequal variances. *Ann. Math. Stat.*, **11**, 475-476 (abstract).

Paper received: February, 1961

SOME LIMIT THEOREMS FOR JOINT DISTRIBUTIONS

By J. SETHURAMAN

Indian Statistical Institute

SUMMARY. In this paper we discuss the convergence of a sequence of joint distributions when it is known that the associated sequences of marginal and conditional distributions converge. Illustrating by means of an example that the joint distributions need not converge weakly when the marginal and conditional distributions converge only weakly, several results are obtained by suitably strengthening the modes of convergence of the latter distributions. An illustrative application of these results is given.

1. INTRODUCTION

When Z is the product space $(X \times Y)$, a random variable ζ on (Z, U) is of the form (ξ, η) where ξ is a random variable on (X, S) and η on (Y, T) . Let $\{\zeta_n\}$ be a sequence of random variables on Z and let $\lambda_n(\cdot)$, $\mu_n(\cdot)$ and $\nu_n(x, \cdot)$ denote the joint distribution of (ξ_n, η_n) , the marginal distribution of ξ_n and the conditional distribution η_n given $\xi_n = x$, respectively. The theorems of the present work are concerned with the convergence of $\{\lambda_n\}$ when $\{\mu_n\}$ and $\{\nu_n\}$ are known to converge in some sense. In an earlier paper this problem has been mentioned by Sukhatme and Sethuraman (1959). Theorem 2 was referred to and used earlier in another paper by the author (Sethuraman, 1961).

2. NOTATIONS AND PRELIMINARIES

Before embarking on the statement and the proofs of our theorems, we explain in this section our notations and mention some well-known results which form the basic tools of this paper.

Throughout this work we will be concerned with two measure spaces (X, S) and (Y, T) , their product $(Z, U) = (X \times Y, S \times T)$ and a sequence of random variables (ξ_n, η_n) taking values in Z . The distribution of (ξ_n, η_n) on (Z, U) will be denoted by λ_n while the distribution of ξ_n on (X, S) will be denoted by μ_n . We assume that the conditional distribution of η_n given $\xi_n = x$ exists as a probability measure, i.e. there exists a function $\nu_n(x, B)$ which is a probability measure on T for each $x \in X$ and is a measurable function on X for each $B \in T$ and further satisfies the equation

$$\lambda_n(A \times B) = \int_A \nu_n(x, B) d\mu_n \text{ for all } A \in S \text{ and } B \in T.$$

In this case, if C is any set in U and C_x for each $x \in X$ denotes the sub-set $\{y : y \in Y, (x, y) \in C\}$, then $\lambda_n(C) = \int_X \nu_n(x, C_x) d\mu_n$. (See Halmos, 1950).

If $p_1, p_2, \dots$, is a sequence of probability measures defined on a measurable space (M, V) , we shall say p_n converges strongly to p ($p_n \rightarrow p$ in symbols) if $p_n(C) \rightarrow p(C)$ for each $C \in V$. It is well known that for a given sequence p_n there exists a p such that $p_n \rightarrow p$ if and only if $\lim p_n(C)$ exists for all $C \in V$; in this case the p_n 's are equicontinuous, i.e. if C_k is any decreasing sequence of sets with $\bigcap_{k=1}^{\infty} C_k = \phi$, the null set, then $\sup_n p_n(C_k) \rightarrow 0$ as $k \rightarrow \infty$. (See Halmos, 1950). It is also known that $p_n \rightarrow p$ if and only if $\int g dp_n \rightarrow \int g dp$ for all bounded measurable functions g . (See Halmos, 1950). If the densities $f_n(m)$ of p_n with respect to some σ -finite measure p_0 , converge in measure $[p_0]$ to a density function $f(m)$ then there is a p such that $p_n \rightarrow p$. (See Scheffe, 1947). The above is a sufficient condition for the strong convergence of a sequence of probability measures $\{p_n\}$ which is convenient in practice.

We shall also require the notion of weak convergence of probability measures. This requires that the basic space be topological, and that all continuous functions be measurable. If $p_1, p_2, \dots$, are probability measures on such a space M , we shall say p_n converges to p weakly ($p_n \Rightarrow p$ in symbols) if $\int g dp_n \rightarrow \int g dp$ for all bounded continuous functions $g(m)$ on M . A set C is said to be a continuity set of p if $p(\text{bd } C) = 0$ where $\text{bd } C$ is the boundary of C . $p_n \Rightarrow p$ if and only if $\lim p_n(C) = p(C)$ for each C that is a continuity set of p . (See Billingsley, 1956). Further, if M is separable complete metric, $\{p_n\}$ is compact under weak convergence if and only if, for each $\epsilon > 0$, there is a compact set $C \subset M$ with $p_n(C) \geq 1 - \epsilon$ for all n . (See Prohorov, 1936 ; Varadarajan, 1958).

For the formulation of one of our theorems we require the notion of UC^* convergence (allied to that of Parzen (1954)) of a family of sequences of probability measures. Let $\nu_n(\theta, \cdot)$, $n = 0, 1, \dots$ be a family of sequences of probability measures on $M = R_k$, the euclidean space of k dimensions. It is assumed that the index θ takes values in a compact metric space I . Let $\phi_n(t, \theta)$ denote the characteristic function of $\nu_n(\theta, \cdot)$, i.e.

$$\phi_n(t, \theta) = \int_M \exp(i t m') \nu_n(\theta, dm), \quad n = 0, 1, \dots \quad \dots (2.1)$$

Definition: $\nu_n(\theta, \cdot)$ is said to converge to $\nu_0(\theta, \cdot)$ in the UC^* sense relative to $\theta \in I$ if

$$(a) \sup_{\theta \in I} |\phi_n(t, \theta) - \phi(t, \theta)| \rightarrow 0 \text{ as } n \rightarrow \infty$$

and

$$\phi_0(t, \theta) \text{ is equicontinuous in } \theta \text{ at } t = 0 \quad \dots (2.2)$$

and

$$(b) \phi_0(t, \theta) \text{ is a continuous function of } \theta \text{ for each } t. \quad \dots (2.3)$$

SOME LIMIT THEOREMS FOR JOINT DISTRIBUTIONS

We mention here that Parzen (1954) uses (2.2) alone as the definition of UC^* convergence. (2.2) and (2.3) imply that $\phi_0(t, \theta)$ is continuous in t and θ . (2.2) and (2.3) also imply the following:

$$\theta_n \rightarrow \theta_0 \text{ implies } \phi_n(t, \theta_n) \rightarrow \phi_0(t, \theta_0) \quad \dots \quad (2.4)$$

which means
$$\nu_n(\theta_n, \cdot) \implies \nu_0(\theta_0, \cdot); \quad \dots \quad (2.5)$$

$$\theta_n \rightarrow \theta_0 \text{ implies } \phi_0(t, \theta_n) \rightarrow \phi_0(t, \theta_0) \quad \dots \quad (2.6)$$

which means
$$\nu_0(\theta_n, \cdot) \implies \nu_0(\theta_0, \cdot). \quad \dots \quad (2.7)$$

3. MAIN THEOREMS

Before proceeding to state and prove the first of our main results we establish the following lemma.

Lemma 1: *Let $\{p_n\}$ be a sequence of probability measures on an arbitrary measure space (M, V) converging strongly to a measure p . Let $\{u_n\}$ be a sequence uniformly bounded V -measurable functions converging almost everywhere $[p]$ to a function u . Then $\int_M u_n dp_n \rightarrow \int_M u dp$.*

Proof: Let $|u(m)| \leq A$ for all n .

Define
$$D_n = \bigcup_{k=n}^{\infty} \{m : |u_k(m) - u(m)| > \epsilon\}.$$

We know that the D_n are decreasing and that if $\bigcap_{n=1}^{\infty} D_n = D$ then $\mu(D) = 0$. We have

$$\begin{aligned} \left| \int_M u_n dp_n - \int_M u dp \right| &= \left| \int_M (u_n - u) dp_n + \int_M u dp_n - \int_M u dp \right| \\ &\leq \left| \int_{D_n} (u_n - u) dp_n \right| + \left| \int_{M-D_n} (u_n - u) dp_n \right| + \left| \int_M u dp_n - \int_M u dp \right| \\ &\leq 2Ap_n(D_n) + \epsilon + \left| \int_M u dp_n - \int_M u dp \right| \dots \quad (3.1) \end{aligned}$$

The first term tends to zero since p_n are equicontinuous and the last term tends to zero since $p_n \rightarrow p$. Since $\epsilon > 0$ is arbitrary, the lemma is proved.

Suppose that, in some sense, the (marginal) distributions μ_n converge to μ and the (conditional) distributions $\nu_n(x, \cdot)$ converge to $\nu(x, \cdot)$. Further, let the joint distributions λ_n converge. It is plausible that λ_n then converges to the distribution λ_0 defined by μ and $\nu(\cdot, \cdot)$, i.e.

$$\lambda_0(A \times B) = \int_A \nu(x, B) d\mu \quad \dots \quad (3.2)$$

over rectangle sets $A \times B$. This defines a distribution λ_0 uniquely on (Z, U) . In what follows, by λ_0 we mean the distribution defined by the relation (3.2).

Theorem 1 : *If the sequence of (marginal) distributions $\{\mu_n\}$ converges strongly to μ and if for almost all $x[\mu]$ the sequence of (conditional) distributions $\{\nu_n(x, \cdot)\}$ converges strongly to $\nu(x, \cdot)$ then the sequence of (joint) distributions $\{\lambda_n\}$ converges strongly to λ_0 .*

Proof: Let $g(x, y)$ be any U -measurable function on Z , bounded by K . We define sequence of S -measurable functions $v_n(x)$ as follows :

$$v_n(x) = \int_Y g(x, y) \nu_n(x, dy). \quad \dots (3.3)$$

It is plain that $|v_n(x)| \leq K$ and that

$$v_n(x) \rightarrow v(x) = \int_Y g(x, y) \nu(x, dy) \quad \dots (3.4)$$

for almost all $x[\mu]$.

On application of Lemma 1 we find that

$$\begin{aligned} \int g(x, y) d\lambda_n &= \int_X d\mu_n \int_Y g(x, y) \nu_n(x, dy) \\ &= \int_X v_n(x) d\mu_n \rightarrow \int_X v(x) d\mu \\ &= \int_X d\mu \int_Y g(x, y) \nu(x, dy) \\ &= \int g(x, y) d\lambda \quad \dots (3.5) \end{aligned}$$

where $\lambda(C) = \int_x \nu(x, C_x) d\mu$ with $C_x = \{y : y \in Y, (x, y) \in C\}$. (See Halmos, 1950).

This λ is the same as λ_0 . Thus $\lambda_n \rightarrow \lambda_0$.

Theorem 2: *If the sequence of (marginal) distributions $\{\mu_n\}$ converges strongly to μ and if the sequence of (conditional) distributions $\{\nu_n(x, \cdot)\}$ converges weakly to $\nu(x, \cdot)$ for almost all $x[\mu]$, then the sequence of (joint) distributions $\{\lambda_n\}$ converges weakly to λ_0 .*

Proof: The proof of this theorem is on the same lines as Theorem 1 and so is omitted.

In Theorems 1 and 2 we assumed that the marginal distributions converge strongly. We now ask ourselves what happens if the marginal distributions converge only weakly. Naturally, one expects that one shall have to strengthen the mode of convergence of the conditional distributions. We have given an example in section 4 to illustrate this fact, that, in general, such strengthening would be necessary. The difficulty in this situation is that the conditional distribution at the n -th stage is defined almost everywhere with respect to μ_n and the $[\mu_n]$ null x -sets, the sets of misbehaviour of $\nu_n(x, \cdot)$, vary with n . Thus we should introduce some smoothness restriction on the conditional distributions.

SOME LIMIT THEOREMS FOR JOINT DISTRIBUTIONS

Before presenting the last of our theorems we prove several lemmas.

In the following lemmas, I is a compact metric space and $M = R_k$. $g(\theta, m)$ is a bounded continuous function on $I \times M$. J is any bounded interval in M . $\{v_n(\theta, \cdot)\}$ $n = 0, 1, \dots$ is a family of sequences of probability measures on M indexed by θ in I . $v_n(\theta, \cdot)$ converges to $v_0(\theta, \cdot)$ the UC^* sense relative to $\theta \in I$. Let

$$v_n(\theta) = \int g(\theta, m) v_n(\theta, dm), \quad n = 0, 1, \dots \quad \dots (3.6)$$

Lemma 2: $g(\theta, m)$ is equicontinuous in θ at each m , i.e. if $\theta_n \rightarrow \theta_0$, then $g(\theta_n, m) \rightarrow g(\theta_0, m)$ uniformly in $m \in J$.

Proof: This follows immediately from the uniform continuity of $g(\theta, m)$ on $I \times J$.

$$\text{Lemma 3: } \sup_{\theta} |v_n(\theta) - v_0(\theta)| \rightarrow 0 \text{ as } n \rightarrow \infty. \quad \dots (3.7)$$

Proof: This lemma is a simple corollary of the general results of Ranga Rao (1960), (1961). We however, give here a short proof for the sake of continuity.

If (3.7) were not true, then there would be a sequence $\{\theta_n\}$ and $\alpha > 0$, such that

$$|v_n(\theta_n) - v_0(\theta_n)| \geq \alpha \text{ for each } n. \quad \dots (3.8)$$

Since I is compact, there is a subsequence $\{\theta_{n_r}\}$ such that $\theta_{n_r} \rightarrow \theta_0$ as $r \rightarrow \infty$.

We then have

$$\begin{aligned} |v_{n_r}(\theta_{n_r}) - v_0(\theta_{n_r})| &= \left| \int g(\theta_{n_r}, m) v_{n_r}(\theta_{n_r}, dm) - \int g(\theta_{n_r}, m) v_0(\theta_{n_r}, dm) \right| \\ &\leq \int |g(\theta_{n_r}, m) - g(\theta_0, m)| v_{n_r}(\theta_{n_r}, dm) \\ &\quad + \left| \int g(\theta_0, m) v_{n_r}(\theta_{n_r}, dm) - \int g(\theta_0, m) v_0(\theta_0, dm) \right| \\ &\quad + \left| \int g(\theta_0, m) v_0(\theta_0, dm) - \int g(\theta_0, m) v_0(\theta_{n_r}, dm) \right| \\ &\quad + \int |g(\theta_0, m) - g(\theta_{n_r}, m)| v_0(\theta_{n_r}, dm). \end{aligned}$$

Given any $\epsilon > 0$ the first, second, third and fourth terms on the right hand side can each be made $< \epsilon$ by using Lemma 2 and (2.5), (2.5), (2.7), and Lemma 2 and (2.7), respectively if $r \geq R$. Since $\epsilon > 0$ is arbitrary, this is a contradiction to (3.8). Hence the lemma.

Lemma 4: $v_0(\theta)$ is continuous in θ .

Proof: This follows immediately from (2.7).

Let $u_n(m)$ be a sequence of functions on a separable complete metric space M converging to a bounded continuous function $u(m)$ uniformly on every compact set. Let μ_n be a sequence of probability measures on M converging weakly to μ . We then have the following lemma.

Lemma 5 : $\int u_n(m) d\mu_n \rightarrow \int u(m) d\mu$.

Proof : The proof is immediate.

In the following theorem $Y = R_k$ and X is any separable complete metric space. Here we impose certain conditions on the conditional distributions similar to those employed by Parzen (1954), Steck (1957) and others.

Theorem 3 : *Let the sequence of (marginal) distributions $\{\mu_n\}$ converge weakly to μ . Let the sequence of (conditional) distributions $\{\nu_n(x, \cdot)\}$ converge to $\nu(x, \cdot)$ in the UC^* sense relative to $x \in I$ for every compact subset I of X . Then the sequence $\{\lambda_n\}$ of (joint) distributions converges weakly to λ_0 .*

Proof : Let $g(x, y)$ be any bounded continuous function on $X \times Y$.

$$\text{Let } u_n(x) = \int_Y g(x, y) \nu_n(x, dy), \quad u(x) = \int_Y g(x, y) \nu(x, dy).$$

Lemma 3 shows that $u_n(x) \rightarrow u(x)$ uniformly in $x \in I$ for every compact $I \subset X$. Lemma 4 shows that $u(x)$ is continuous in x .

Now, using Lemma 5,

$$\begin{aligned} \int g(x, y) d\lambda_n &= \int_X d\mu_n \int_Y g(x, y) \nu_n(x, dy) \\ &= \int u_n(x) d\mu_n \rightarrow \int u(x) d\mu \\ &= \int_X d\mu \int_Y g(x, y) \nu(x, dy) \\ &= \int g(x, y) d\lambda \end{aligned}$$

where $\lambda(C) = \int_X d\mu \int_{C_x} \nu(x, dy)$ with $C_x = \{y : (x, y) \in C\}$. (See Halmos, 1950). Thus

$\lambda = \lambda_0$ and $\lambda_n \Rightarrow \lambda_0$.

4. A COUNTER EXAMPLE

We present below an example to show that some such conditions, as imposed in Theorems 1, 2 and 3 are, in general, necessary.

X and Y are the real line and S and T the usual field of Borel sets. The random variable (ξ_n, η_n) takes the values $(1/n, 1/n)$, $(1/n, 1+1/n)$, $(1+1/n, 1/n)$ and $(1+1/n, 1+1/n)$ with probabilities $\frac{1}{8}$, $\frac{3}{8}$, $\frac{3}{8}$, and $\frac{1}{8}$ respectively if n is even and with probabilities $\frac{3}{8}$, $\frac{1}{8}$, $\frac{1}{8}$ and $\frac{3}{8}$ respectively if n is odd.

SOME LIMIT THEOREMS FOR JOINT DISTRIBUTIONS

It is easy to see that the marginal distributions of ξ_n and η_n converge weakly to the same distribution with masses $\frac{1}{2}$ and $\frac{1}{2}$ at 0 and 1. The conditional distributions are trivially convergent. The joint distributions do not converge.

5. AN ILLUSTRATIVE APPLICATION

In section 1 we have already mentioned some applications of the results of section 3 made in earlier works. As an illustration we can deduce the asymptotic distribution of several sample quantiles from that of a single quantile and Theorem 2. In particular, assuming the theorem for the asymptotic distribution of the sample median (see Smirnov, 1949; Cramer, 1946), we will show that the sample first quartile and the median are jointly asymptotically normally distributed.

Let $z_1, \dots, z_{4n+3}$ be $4n+3$ independent observations on a random variable Z with distribution function $F(z)$ which possesses a density function $f(z)$. It is assumed that $f(z)$ is continuous and nonzero at θ , the population median and at δ , the population first quartile.

Let us denote $\sqrt{n}(z_{(2n+2)} - \theta)$, $\sqrt{n}(z_{(n+1)} - \delta)$ by (ξ_n, η_n) where $z_{(2n+2)}$ is the sample median and $z_{(n+1)}$ is the sample first quartile. When ξ_n is fixed at x , η_n is the normalised median of a sample of size $(2n+1)$ on the random variable Z truncated to the region $\left(-\infty, \theta + \frac{x}{\sqrt{n}}\right)$, and hence is asymptotically normal. Some algebraic computations show that the mean and variance of the limiting conditional distribution are $\frac{x f(\theta)}{2 f(\delta)}$ and $\frac{1}{32 f^2(\delta)}$. Since the densities converge pointwise (Cramer, 1946)

ξ_n tends strongly to the normal distribution with mean zero and variance $\frac{1}{16 f^2(\theta)}$.

An application of Theorem 3 shows that the joint distribution of (ξ_n, η_n) converges to the bivariate normal distribution with means zero, zero, variances

$\frac{1}{16 f^2(\theta)}$, $\frac{1}{64 f^2(\delta)}$ and correlation $\frac{1}{\sqrt{3}}$.

REFERENCES

- BILLENGSLEY, P. (1956): The invariance principle for dependent random variables. *Trans. Amer. Math. Soc.*, **83**, 250-268.
- CRAMÉR, H. (1946): *Mathematical Methods of Statistics*, Princeton University Press.
- HALMOS, P. R. (1950): *Measure Theory*, Van Nostrand Company, New York.
- PARZEN, E. (1954): *On Uniform Convergence of Families of Sequences of Random Variables*, University of California Publications in Statistics, **2**, No. 2, 23-54.
- PROHOROV, YU. V. (1956): *Convergence of Random Processes and Limit Theorems in the Theory of Probability*, *Teoriya Veroyat i. Primen*, **1**, 177-238 (Russian).

SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS : SERIES A

- RANGA RAO, R. (1960): Some limit theorems in probability theory. Unpublished D. Phil. thesis, submitted to the University of Calcutta.
- (1961): Some theorems on weak convergence of measures and its applications, submitted to *Ann. Math. Stat.*
- SCHEFFÉ, H. (1947): A useful convergence theorem for probability distributions. *Ann. Math. Stat.*, **18**, 434-438.
- SETHURAMAN, J. (1961): Some limit theorems connected with fractile graphical analysis. *Sankhyā*, Series A, **23**, 79-90.
- SETHURAMAN, J. and SUKHATME, B. V. (1959): Joint asymptotic distribution of U -statistics and order statistics. *Sankhyā*, **21**, 289-298.
- SMIRNOV, N. V. (1949): Limit distributions for the terms of a variational series. *Trudy. Mat. Inst. A. Steklov.*, **25**, 1-61 (Russian).
- STECK, G. P. (1957): *Limit Distributions for Conditional Distributions*, University of California Publications in Statistics, **2**, No. 12, 237-284.
- VARADARAJAN, V. S. (1958): Convergence of stochastic processes. Unpublished D. Phil. thesis, submitted to the University of Calcutta.

Paper received : June, 1961.

BEST ONE-SIDED BOUNDS FOR INFINITELY DIVISIBLE RANDOM VARIABLES

By HOWARD G. TUCKER

University of California, Riverside, USA

SUMMARY. The essential supremum (essential infimum) is found for a random variable with infinitely divisible distribution which is bounded above (below). Necessary and sufficient conditions are given that almost all sample functions of a continuous separable stochastic process with independent increments be discrete. A direct probabilistic proof is given that a random variable with infinitely divisible distribution be unbounded.

1. INTRODUCTION

A random variable X is said to be infinitely divisible if its characteristic function $\phi_X(u)$ can be written

$$\phi_X(u) = \exp \left\{ iu\gamma + \int_{-\infty}^{\infty} \left(e^{iux} - 1 - \frac{iux}{1+x^2} \right) \frac{1+x^2}{x^2} dG(x) \right\}, \quad \dots \quad (1.1)$$

where γ is a fixed real constant and G is a bounded, real-valued, non-decreasing function. Clearly, X is degenerate (i.e., constant with probability one) if and only if G is identically constant. Pakshirajan and Chatterjee (1956) proved that a non-degenerate infinitely divisible random variable X is unbounded. Recently, Baxter and Shapiro (1960) found necessary and sufficient conditions that a random variable with an infinitely divisible distribution be bounded above (below), and they exhibited an upper (lower) bound. This result was used to obtain necessary and sufficient conditions that almost all of the sample functions of a separable process with stationary, independent increments be non-decreasing (non-increasing).

The purpose of this note is to sharpen both of these results. In Section 2 it will be shown (Theorem 1) that the one-sided bounds exhibited by Baxter and Shapiro (1960) are best bounds. In Section 3 necessary and sufficient conditions are given that almost all of the sample functions of a separable, continuous, decomposable process be discrete, i.e., almost every sample function be a function of bounded variation over every bounded interval and such that its difference over every interval be the sum of its discontinuities. In Section 4 a simple and direct proof of the unboundedness of a non-degenerate infinitely divisible random variable is given.

2. THE BEST ONE-SIDED BOUNDS

The technique used in this section and in Section 4 consists of a particular interpretation of a certain class of probability distributions. We first discuss such distributions.

Let $X_1, X_2, \dots$, be a sequence of independent, identically distributed random variables with common distribution function F . Let Y be a random variable with a

Poisson distribution, $EY = \lambda > 0$, and assume that $Y, X_1, X_2, \dots, X_n, \dots$ are independent. Then the distribution of the random sum of random variables

$$Z = X_1 + X_2 + \dots + X_Y$$

has as characteristic function

$$\phi_Z(u) = \exp \lambda \int_{-\infty}^{\infty} (e^{iux} - 1) dF(x).$$

Conversely, if G is a bounded, non-decreasing function over $(-\infty, \infty)$, and if a random variable Z has as its characteristic function

$$\phi_Z(u) = \exp \int_{-\infty}^{\infty} (e^{iux} - 1) dG(x),$$

then Z has the same distribution as a random sum,

$$U_1 + \dots + U_V,$$

where $\{U_n\}$ are independent, identically distributed random variables with common distribution function

$$F(x) = G(x)/(G(+\infty) - G(-\infty)),$$

where V has a Poisson distribution with $EV = G(+\infty) - G(-\infty)$, and $V, U_1, U_2, \dots$ are independent.

Baxter and Shapiro (1960) proved that if X is a random variable with infinitely divisible distribution (1.1) and which is bounded above, then

$$P[X \leq \gamma - \int_{-\infty}^{-0} \frac{1}{x} dG(x)] = 1,$$

the existence of the above integral being a consequence of the hypothesis of X being bounded above. We now sharpen this result.

Theorem 1: *If X is a random variable with infinitely divisible distribution (1.1) and is with probability one bounded above by a constant, then*

$$\text{ess sup } X = \gamma - \int_{-\infty}^{-0} \frac{1}{x} dG(x).$$

(ii) *If X is a random variable with infinitely divisible distribution (1.1) and is bounded below by a constant with probability one, then*

$$\text{ess inf } X = \gamma - \int_{0+}^{\infty} \frac{1}{x} dG(x).$$

BEST ONE-SIDED BOUNDS FOR INFINITELY DIVISIBLE RANDOM VARIABLES

Proof: We need present only the proof to part (i) of the theorem. By Theorem 1 in Baxter and Shapiro (1960) the following conditions hold :

- (i) $\int_u^\infty \frac{1+x^2}{x^2} dG(x) = 0$ for all $u > 0$,
- (ii) $\sigma^2 = G(+0) - G(-0) = 0$, and ...
- (iii) $\lim_{\varepsilon \downarrow 0} \int_{-1}^{-\varepsilon} \int_{-\infty}^u \frac{1+x^2}{x^2} dG(x) du < \infty$.

We denote $M(u) = \int_{-\infty}^u \frac{1+x^2}{x^2} dG(x)$ for $u < 0$. In case $M(-0) = 0$, the assertion of the theorem is trivial. We first consider the special case where $0 < M(-0) < \infty$. As was mentioned earlier the assumption of boundedness above implies the existence of the integral $\int_{-\infty}^{-0} \frac{1}{x} dG(x)$, and thus we may write

$$\log \phi_X(u) = \int_{-\infty}^{-0} (e^{iux} - 1) dM(x) + iuA_0,$$

where

$$A_0 = \gamma - \int_{-\infty}^{-0} \frac{1}{x} dG(x).$$

Let Y be a random variable whose characteristic function is given by

$$\log \phi_Y(u) = \int_{-\infty}^{-0} (e^{iux} - 1) dM(x).$$

Let $\{U_n\}$ be a sequence of independent identically distributed random variables with common distribution function

$$F(x) = \begin{cases} M(x)/M(-0) & \text{if } x < 0 \\ 1 & \text{if } x \geq 0. \end{cases}$$

Let V be a random variable with Poisson distribution, $EV = M(-0)$, and such that $V, U_1, \dots, U_n, \dots$ are independent. Then the distribution of Y is the same as that of $U_1 + \dots + U_V$. Since $P[U_n < 0] = 1$ for all n , then it follows that $P[Y \leq 0] = 1$, and

$$P[Y = 0] = e^{-M(-0)} > 0.$$

Hence $\text{ess sup } Y = 0$, from which we obtain that $\text{ess sup } X = A_0$, thus proving the theorem in the case when $M(-0) < \infty$. Now we prove it in the case $M(-0) = \infty$.

Since from Baxter and Shapiro (1960) we know that $\text{ess sup } X \leq A_0$, then we need only show that for every $\epsilon > 0$ that

$$\text{ess sup } X > \gamma - \int_{-\infty}^{-\epsilon} \frac{x}{1+x^2} dM(x).$$

If we let

$$S_\epsilon = X - \gamma + \int_{-\infty}^{-\epsilon} \frac{x}{1+x^2} dM(x),$$

then it is sufficient to prove that $P[S_\epsilon > 0] > 0$ for every $\epsilon > 0$. We first observe that

$$\log \phi_{S_\epsilon}(u) = \int_{-\infty}^{-\epsilon} (e^{iux} - 1) dM(x) - iu \int_{-\epsilon}^{-\epsilon} \frac{x}{1+x^2} dM(x). \text{ Let } K_\epsilon = - \int_{-\epsilon}^{-\epsilon} \frac{x}{1+x^2} dM(x).$$

Since $M(-0) = \infty$, then $K_\epsilon > 0$ for every $\epsilon > 0$. Let $\eta > 0$ be such that $-\eta$ is a continuity point of $M(x)$, and let U_η and V_η be independent random variables such that

$$\log \phi_{U_\eta}(u) = \int_{-\infty}^{-\eta} (e^{iux} - 1) dM(x),$$

$$\log \phi_{V_\eta}(u) = \int_{-\eta}^{-\epsilon} (e^{iux} - 1) dM(x).$$

Now the distribution of S_ϵ is the same as that of $U_\eta + V_\eta + K_\epsilon$. But $V_\eta \rightarrow 0$ in probability as $\eta \rightarrow 0$, and $V_\eta + K_\epsilon \rightarrow K_\epsilon > 0$ in probability as $\eta \rightarrow 0$. Hence there exists an $\eta > 0$ sufficiently small so that $P[V_\eta + K_\epsilon > 0] > 0$. Since

$$[V_\eta + K_\epsilon > 0][U_\eta = 0] \subset [S_\epsilon > 0],$$

and since (by the special case proved when $M(-0) < \infty$)

$$P[U_\eta = 0] = e^{-M(-\eta)} > 0,$$

we obtain

$$0 < P[V_\eta + K_\epsilon > 0]P[U_\eta = 0] \leq P[S_\epsilon > 0],$$

which proves the theorem.

3. DISCRETENESS OF SAMPLE FUNCTIONS OF A SEPARABLE CONTINUOUS DECOMPOSABLE PROCESS

Baxter and Shapiro (1960) found necessary and sufficient conditions that there exists a constant A such that almost all the sample functions of $X(t) - At$ be non-decreasing over $[0, 1]$, where $X(t)$ is a separable process with stationary independent increments. In this section, this result is extended to a process $X(t)$ which is a separable continuous decomposable process (see Loève (1960), pp. 534-554). At the same

BEST ONE-SIDED BOUNDS FOR INFINITELY DIVISIBLE RANDOM VARIABLES

time it is sharpened in that necessary and sufficient conditions are given that almost all sample functions over a t -interval $[0, T]$ are discrete.

We shall say that a real-valued function g defined over an interval $[a, b]$ is discrete if and only if (i) it is of bounded variation, and (ii) for every $x \in [a, b]$, $g(x)$ equals the sum of the discontinuities of g in $[a, x]$. A continuous decomposable process is a process with independent increments and is continuous at each value of t with probability 1. At each t , the characteristic function of $X(t)$ is given by

$$\log \phi_{X(t)}(u) = iu\alpha(t) + \int_{-\infty}^{\infty} \left(e^{iux} - 1 - \frac{iux}{1+x^2} \right) \frac{1+x^2}{x^2} dG_t(x), \quad \dots (3.1)$$

where $\alpha(t)$ is a continuous function of t , $G_t(x)$ is for each t a bounded, non-decreasing function of x and for each x is a non-decreasing, continuous function of t . (Note: in what follows, $G_1(x) = G_t(x)$ at $t = 1$).

Theorem 2: *Let $X(t)$ be a separable, continuous, decomposable process, $t \in [0, 1]$. In order that there exists a continuous (sure) function f over $[0, 1]$ such that almost all sample functions of $X(t) - f(t)$ are discrete (as defined above) it is necessary and sufficient that*

$$(i) \quad G_1(+0) - G_1(-0) = 0,$$

and

$$(ii) \quad \lim_{\varepsilon \downarrow 0} \left(\int_{\varepsilon}^1 \int_u^{\infty} \frac{1+x^2}{x^2} dG_1(x) du + \int_{-1}^{-\varepsilon} \int_{-\infty}^u \frac{1+x^2}{x^2} dG_1(x) du \right) < \infty.$$

In such a case,

$$f(t) = \alpha(t) - \int_{-\infty}^{\infty} \frac{1}{x} dG_t(x).$$

Proof: We first prove that conditions (i) and (ii) are sufficient. First, we may assume that $P[X(0) = 0] = 1$. We shall use the fact that, for $0 \leq t_1 < t_2$,

$$\int_u^{\infty} \frac{1+x^2}{x^2} dG_{t_2}(x) - \int_u^{\infty} \frac{1+x^2}{x^2} dG_{t_1}(x)$$

is the (finite) expected number of jumps of $X(t)$ of magnitude $\geq u > 0$ during the interval $[t_1, t_2]$, and

$$\int_{-\infty}^u \frac{1+x^2}{x^2} dG_{t_2}(x) - \int_{-\infty}^u \frac{1+x^2}{x^2} dG_{t_1}(x)$$

is the expected number of jumps of $X(t)$ of magnitude $\leq u < 0$ during $[t_1, t_2]$. Because the number of jumps of any magnitude outside a neighbourhood of 0 is a

non-negative random variable we obtain (from the additivity property of expectation) that if $0 < t < 1$, then

$$\int_u^\infty \frac{1+x^2}{x^2} dG_t(x) \leq \int_u^\infty \frac{1+x^2}{x^2} dG_1(x) \text{ for } u > 0,$$

and

$$\int_{-\infty}^u \frac{1+x^2}{x^2} dG_t(x) \leq \int_{-\infty}^u \frac{1+x^2}{x^2} dG_1(x) \text{ for } u > 0.$$

Thus, condition (ii) implies that

$$\lim_{\varepsilon \downarrow 0} \left(\int_\varepsilon^1 \int_u^\infty \frac{1+x^2}{x^2} dG_t(x) du + \int_{-1}^{-\varepsilon} \int_{-\infty}^u \frac{1+x^2}{x^2} dG_t(x) du \right) < \infty \quad \dots (3.2)$$

for all $t \in [0, 1]$.

Now let $U(t)$, $V(t)$ be two independent separable continuous decomposable processes such that

$$\log \phi_{U(t)}(u) = \int_{+0}^\infty \left(e^{iux} - 1 - \frac{iux}{1+x^2} \right) \frac{1+x^2}{x^2} dG_t(x),$$

and

$$\log \phi_{V(t)}(u) = \int_{-\infty}^{-0} \left(e^{ixu} - 1 - \frac{iux}{1+x^2} \right) \frac{1+x^2}{x^2} dG_t(x).$$

Because of condition (i) we may (effectively) write

$$X(t) = \alpha(t) + U(t) + V(t).$$

By (3.2) and integration by parts we have

$$\begin{aligned} & \lim_{\varepsilon \downarrow 0} \int_\varepsilon^1 \int_u^\infty \frac{1+x^2}{x^2} dG_t(x) du \\ &= \lim_{\varepsilon \downarrow 0} \left(\int_1^\infty \frac{1+x^2}{x^2} dG_t(x) - \varepsilon \int_\varepsilon^\infty \frac{1+x^2}{x^2} dG_t(x) - \int_\varepsilon^1 x \frac{1+x^2}{x^2} dG_t(x) \right) < \infty. \quad \dots (3.3) \end{aligned}$$

The existence of the improper integral $\int_{0+}^a f(x)dx$, when $f(x) \rightarrow \infty$ as $x \downarrow 0$ implies that $\varepsilon f(\varepsilon) \rightarrow 0$ as $\varepsilon \downarrow 0$. Hence the hypothesis (ii) implies that

$$\lim_{\varepsilon \downarrow 0} \varepsilon \int_\varepsilon^\infty \frac{1+x^2}{x^2} dG_t(x) = 0.$$

Thus (3.3) implies that

$$\lim_{\varepsilon \downarrow 0} \int_\varepsilon^\infty \frac{1+x^2}{x} dG_t(x) = \int_{0+}^\infty \frac{1+x^2}{x} dG_t(x) < \infty,$$

BEST ONE-SIDED BOUNDS FOR INFINITELY DIVISIBLE RANDOM VARIABLES
and therefore

$$\int_{0+}^{\infty} \frac{1}{x} dG_t(x) < \infty \quad \text{for all } t \in [0, 1].$$

In a similar manner, (3.2) implies that $\int_{-\infty}^{0-} \frac{1}{x} dG_t(x) > -\infty$ for all $t \in [0, 1]$.

Hence

$$U'(t) = U(t) + \int_{0+}^{\infty} \frac{1}{x} dG_t(x)$$

is a process whose characteristic function is given by

$$\log \phi_{U'(t)}(u) = \int_{0+}^{\infty} \left(e^{iux} - 1 \right) \frac{1+x^2}{x^2} dG_t(x).$$

By Theorem 1, if $0 \leq t_1 < t_2 \leq 1$, $U'(t_2) - U'(t_1)$ is bounded below by zero, and therefore almost all of the sample functions of $U'(t)$ are non-decreasing. For fixed t , $\epsilon > 0$, let $U^{(\epsilon)}(t)$ denote the sum of the jumps of $U'(\cdot)$ (or of $X(\cdot)$) during the interval $[0, t]$ of magnitude not less than ϵ . Since

$$\log \phi_{U^{(\epsilon)}(t)}(u) = \int_{\epsilon}^{\infty} \left(e^{iux} - 1 \right) \frac{1+x^2}{x^2} dG_t(x),$$

the distribution of $U^{(\epsilon)}(t)$ converges to that of $U'(t)$ as $\epsilon \downarrow 0$. Since, for fixed t , $U^{(\epsilon)}(t)$ is monotone non-decreasing as $\epsilon \downarrow 0$, we obtain

$$U'(t) = \lim_{\epsilon \downarrow 0} U^{(\epsilon)}(t)$$

with probability one. Hence $U'(t)$ equals the sum of all positive jumps over $[0, t]$. Similarly, if we define

$$V_t = V_t + \int_{-\infty}^{0-} \frac{1}{x} dG_t(x),$$

we obtain that V_t is the sum of all the negative jumps during $[0, t]$. Now it is known that almost all of the sample functions of $X(t)$ have discontinuities only of the first kind. Hence by condition (i) we may now conclude that

$$X(t) - U'(t) - V'(t) = \alpha(t) - \int_{-\infty}^{\infty} \frac{1}{x} dG_t(x)$$

is a continuous sure function, and the sufficiency is proved.

Conversely, let us suppose that there exists a continuous function f over $[0, 1]$ such that $X(t) - f(t)$ is a random function for which almost all sample functions are

discrete. Let $Y(t)$ denote the sum of all positive jumps in $[0, t]$, and let $Z(t)$ denote the sum of all negative jumps in $[0, t]$. If $Y_\epsilon(t)$ denotes the sum of those positive jumps in $[0, t]$ which are in magnitude not less than $\epsilon > 0$, then

$$\log \phi_{Y_\epsilon(t)} = \int_{\epsilon}^{\infty} (e^{iux} - 1) \frac{1+x^2}{x^2} dG_t(x).$$

By hypothesis, $Y_\epsilon(t) \rightarrow Y(t)$ as $\epsilon \downarrow 0$ with probability one for each t . Hence,

$$\log \phi_{Y(t)}(u) = \int_{0+}^{\infty} (e^{iux} - 1) \frac{1+x^2}{x^2} dG_t(x).$$

For each $\epsilon > 0$, and each t , $Y_\epsilon(t)$ has an infinitely divisible distribution (1.1) with

$$\gamma = \int_{\epsilon}^{\infty} \frac{1}{x} dG_t(x),$$

$$G(x) = \begin{cases} G_t(x) & \text{if } x \geq \epsilon \\ G_t(\epsilon) & \text{if } x < \epsilon. \end{cases}$$

Since $Y_\epsilon(t) \rightarrow Y_t$ as $\epsilon \downarrow 0$ with probability one, then by the well-known convergence theorem (Loève (1960), page 300), it follows that $Y(t)$ has an infinitely divisible distribution (1.1) with

$$\gamma = \int_{0+}^{\infty} \frac{1}{x} dG_t(x),$$

$$G(x) = \begin{cases} G_t(x) & \text{if } x > 0 \\ G_t(0+) & \text{if } x \leq 0. \end{cases}$$

Since $Y(1)$ is a non-negative infinitely divisible random variable it is bounded below by zero, and hence by Theorem 2 in Baxter and Shapiro (1960) we obtain

$$\lim_{\epsilon \downarrow 0} \int_{\epsilon}^1 \int_u^{\infty} \frac{1+x^2}{x^2} dG_1(x) du < \infty.$$

Similar consideration for $Z(t)$ implies that

$$\lim_{\epsilon \downarrow 0} \int_{-1}^{-\epsilon} \int_{-\infty}^u \frac{1+x^2}{x^2} dG_1(x) du < \infty.$$

We now prove that $G_1(0+) - G_1(0-) = 0$. Since $X(t) = f(t) + Y(t) + Z(t)$, then

$$\log \phi_{X(t)}(u) = iuf(t) + \int_{-\infty}^{\infty} (e^{iux} - 1) \frac{1+x^2}{x^2} dG_t(x). \text{ If we compare this with (3.1)}$$

we obtain $G_t(0+) - G_t(0-) = 0$ for all $t \in [0, 1]$. Thus conditions (i) and (ii) are necessary.

BEST ONE-SIDED BOUNDS FOR INFINITELY DIVISIBLE RANDOM VARIABLES

4. DIRECT PROOF OF UNBOUNDEDNESS OF INFINITELY DIVISIBLE RANDOM VARIABLES

The proof of the unboundedness of a non-degenerate infinitely divisible random variable given by Pakshirajan and Chatterjee (1956) is essentially a non-probabilistic proof. The proof given by Baxter and Shapiro (1960) is a corollary to their more comprehensive Theorems 1 and 2. Actually, this result is a direct consequence of two rather simple lemmas.

Lemma 1 : *If X and Y are two independent random variables, and if Y is unbounded above (below), then $X+Y$ is unbounded above (below).*

Proof : Suppose Y is unbounded above, and let z be any real number. There does exist an x such that $P[X > x] > 0$. Since Y is unbounded above, $P[Y > z-x] > 0$. Now,

$$[X > x][Y > z-x] \subset [X+Y > z],$$

and because of independence of X and Y we obtain

$$P[X+Y > z] \geq P[X > x] P[Y > z-x] > 0,$$

which proves the lemma.

Lemma 2 : *Let F be a non-decreasing function over $[a, b]$ with $F(b) > F(a)$, and let X be a random variable with characteristic function given by*

$$\log \phi_X(u) = \int_a^b (e^{iux} - 1) dF(x). \text{ If } b < 0, \text{ then } X \text{ is unbounded below, and if}$$

$a > 0$, then X is unbounded above.

Proof : We prove this in the case $a > 0$ only. By the discussion in Section 2 prior to the statement of Theorem 1 it follows that

$$P[X \geq na] = e^{-(F(b)-F(a))} (F(b)-F(a))^n / n! > 0$$

for every positive integer n , which proves the lemma.

From Lemmas 1 and 2 the proof of the unboundedness of an infinitely divisible random variable Z with characteristic function (1.1) is immediate. Since G is not constant then (i) $G(0+) - G(0-) > 0$ or (ii) there is an interval $[a, b]$ such that $G(b) > G(a)$, where $a > 0$ or $b < 0$. If (i) is true, then we may write (from a distributional point of view)

$$Z = U + V,$$

where U and V are independent, and

$$\log \phi_U(u) = -(G(0+) - G(0-)) u^2 / 2,$$

$$\log \phi_V(u) = i\gamma u + \int_{-\infty}^{\infty} \left(e^{iux} - 1 - \frac{iux}{1+x^2} \right) \frac{1+x^2}{x^2} dG(x).$$

Since U has a normal distribution, it is unbounded, and by Lemma 1, Z is unbounded. If (ii) is true, then we may write

$$Z = S + T,$$

where S and T are independent random variables, and

$$\log \phi_S(u) = \int_a^b (e^{iux} - 1) \frac{1+x^2}{x^2} dG(x),$$

$$\log \phi_T(u) = iu \left(\gamma - \int_a^b \frac{1}{x} dG(x) \right)$$

$$+ \int_{-\infty}^a + \int_b^{\infty} \left(e^{iux} - 1 - \frac{iux}{1+x^2} \right) \frac{1+x^2}{x^2} dG(x).$$

By Lemma 2, S is unbounded, and by Lemma 1, Z is therefore unbounded.

REFERENCES

- BAXTER, GLEN and SHAPIRO, J. M. (1960): On bounded infinitely divisible random variables. *Sankhyā*, **22**, 253-260.
- LOËVE, MICHEL (1960): *Probability Theory*, D. Van Nostrand, Princeton, (second edition).
- PAKSHIRAJAN, R. P. and CHATTERJEE, S. D. (1956): On the unboundedness of infinitely divisible laws. *Sankhyā*, **17**, 349-350.

Paper received : June, 1961.

TCHEBYCHEFF-TYPE INEQUALITIES IN TERMS OF THE MEAN DEVIATION

By GERALD J. GLASSER

New York University

SUMMARY. In this paper, inequalities for the frequency of observations lying in an interval are obtained in terms of the mean and deviation similar to the classical Tchebycheff inequalities. The performance of these inequalities are compared with the usual inequalities in some interesting special cases.

1. INTRODUCTION

Given the values of the mean (μ) and the standard deviation (σ) of a population, the use of the well-known Bienaymé-Tchebycheff inequality suggests distribution-free limits for the relative frequency of the observations that lie within a central interval. This particular result, however, is a special case of a general distribution-free inequality in terms of the mean and the standard deviation, that covers any interval including the mean value. This is due to Selberg (1940) who showed

$$P(\mu - k_1\sigma, \mu + k_2\sigma) \geq \begin{cases} 0 & \text{if } k_1k_2 \leq 1 \\ 4 \frac{k_1k_2 - 1}{(k_1 + k_2)^2} & \text{if } k_1k_2 \geq 1 \geq k_2(k_1 - k_2)/2 \\ k_2^2/(1 + k_2^2) & \text{if } k_2(k_1 - k_2)/2 \geq 1 \end{cases} \quad \dots \quad (1.1)$$

where $k_1 \geq k_2 \geq 0$ (the subscripts may be interchanged for $k_2 \geq k_1 \geq 0$) and where $P(a, b)$ is the probability that a variate, x , satisfies $a < x < b$. Two special cases of (1.1) have $k_1 = k_2$, and $k_1 \rightarrow \infty$ for

$$P(\mu - k\sigma, \mu + k\sigma) \geq \begin{cases} 0 & \text{if } k \leq 1 \\ 1 - 1/k^2 & \text{if } k \geq 1 \end{cases} \quad \dots \quad (1.2)$$

(the Bienaymé-Tchebycheff inequality), and

$$P(-\infty, \mu + k\sigma) \geq 1 - 1/(1 + k^2) \quad \text{if } k \geq 0 \quad \dots \quad (1.3)$$

respectively. All of these results can be somewhat improved with additional information about the population, a minimum or maximum value, or some higher order moment, for example (cf. Godwin, 1955). For the information given, however, these are the best possible inequalities.

This note considers comparable results in terms of the average deviation about the mean, δ , and contrasts them with those available in terms of the standard deviation. The use of δ rather than σ may permit a somewhat sharper inequality, particularly when δ is only a small fraction of σ because the latter is greatly inflated by extreme values.

2. RESULTS IN TERMS OF δ AND COMPARISONS

Given δ rather than σ , the best possible result comparable to (1.1) shows

$$P(\mu - t_1\delta, \mu + t_2\delta) \geq \begin{cases} 0 & t_1t_2 \leq (t_1+t_2)/2 \\ 1 - \frac{1}{2} \left(\frac{1}{t_1} + \frac{1}{t_2} \right) & t_1t_2 \geq (t_1+t_2)/2 \end{cases} \quad \dots (2.1)$$

the proof of which follows from $\delta \geq qt_1\delta + pt_2\delta + |qt_1\delta - pt_2\delta|$ where $p = 1 - P(-\infty, \mu + t_2\delta)$ and $q = 1 - P(\mu - t_1\delta, \infty)$. For the special cases $t_1 = t_2$ and $t_1 \rightarrow \infty$

$$P(\mu - t\delta, \mu + t\delta) \geq \begin{cases} 0 & t \leq 1 \\ 1 - 1/t & t \geq 1 \end{cases} \quad \dots (2.2)$$

and

$$P(-\infty, \mu + t\delta) \geq \begin{cases} 0 & t \leq 1/2 \\ 1 - 1/2t & t \geq 1/2. \end{cases} \quad \dots (2.3)$$

The two sets of inequalities may be contrasted by considering a given interval, $\mu - k_1\sigma = \mu - t_1\delta$ to $\mu + k_2\sigma = \mu + t_2\delta$, and comparing the lowest bounds provided by (1.1) and (2.1). The larger lowest bound is obviously the better, and the comparison shows that this is provided by (2.1) if any of the following conditions is satisfied

$$\frac{\delta}{\sigma} < \begin{cases} \frac{2k_1k_2}{k_1+k_2} & k_1k_2 \leq 1 \\ \frac{2k_1k_2[(k_1-k_2)^2+4]}{(k_1+k_2)^3} & k_1k_2 \geq 1 \geq k_2(k_1-k_2)/2 \\ \frac{2k_1k_2}{(k_1+k_2)(1+k_2^2)} & k_2(k_1-k_2)/2 \geq 1. \end{cases} \quad \dots (2.4)$$

For the special case $k_1 = k_2$ the condition for (2.2) to be sharper is

$$\frac{\delta}{\sigma} < k < \frac{\sigma}{\delta} \quad \dots (2.5)$$

while with $k_1 \rightarrow \infty$, (2.3) is better than (1.3) if

$$\frac{\delta}{\sigma} < \frac{2k}{1+k^2} \quad \dots (2.6)$$

These inequalities can also be put in terms of t rather than k , but in any event clearly depend on δ/σ being smaller than some function of the width of the specified interval.

For some intervals, inequalities in terms of δ and σ can be developed that are better than either (1.1) or (2.1), but these will not be considered here.

For a second comparison of (1.1) and (2.1), it may be noted that the inequalities given also permit one to locate, within some interval about the mean, the p -th centile

TCHEBYCHEFF-TYPE INEQUALITIES IN TERMS OF THE MEAN DEVIATION of a population, ζ_p . Thus, (1.3) applied back and forth shows that (cf. Hoeffding, 1952, p. 173).

$$\mu - \sqrt{(1-p)/p} \sigma \leq \zeta_p \leq \mu + \sqrt{p/(1-p)} \sigma \quad \dots \quad (2.7)$$

while (2.3) may be similarly applied to show

$$\mu - \delta/2p \leq \zeta_p \leq \mu + \delta/2(1-p). \quad \dots \quad (2.8)$$

The width of the first interval is $\sigma/\sqrt{p(1-p)}$, while the width of the second is $\delta/2p(1-p)$.

Therefore, (2.8) is sharper if

$$\frac{\delta}{\sigma} < 2\sqrt{p(1-p)}. \quad \dots \quad (2.9)$$

In particular, the inequalities locate the median within one σ , and within one δ , of the mean, and since usually δ will be somewhat smaller than σ , (2.8) provides somewhat narrower limits.

3. VALUES OF δ/σ

The relative precision of (1.1) and (2.1) in the two types of problems discussed depends upon the value of δ/σ , which will vary from one type of population to another. From their definitions it follows that

$$\sigma^2 - \delta^2 = \text{var } |x - \mu|$$

so that their difference depends on the variation in the absolute deviations of the observations from their mean. Thus, for populations with extreme deviations, σ will tend to be somewhat higher than δ and δ/σ will tend to be small.

A distribution-free result due to Chakravarti (1947-48) is of interest in this connection. He shows that for N variates ($N \geq 2$)

$$\left. \begin{aligned} \sqrt{2/N} &\leq \delta/\sigma \leq \sqrt{1-1/N^2} & (N \text{ odd}, \sigma \neq 0) \\ \sqrt{2/N} &\leq \delta/\sigma \leq 1 & (N \text{ even}, \sigma \neq 0) \end{aligned} \right\} \quad \dots \quad (3.1)$$

so, for $N \rightarrow \infty$, δ/σ may range from 0 to 1.

Although the results given are distribution-free, it is somewhat instructive to study their precision in the two problems discussed for various population models, though for such models exact answers, or at least more precise inequalities, may usually be derived.

Thus, for a normal population $\delta/\sigma = \sqrt{2/\pi} = .7979$. This suggests, for example, that (2.5) is satisfied for $.8 \leq k \leq 1.25$ and that (2.6) holds if $.5 \leq k \leq 2.0$. Further, (2.9) is satisfied for $.2 \leq p \leq .8$. So for a normal value of δ/σ , the inequalities based on δ are sharper for locating probabilities or percentiles in the central part of a distribution. For the exponential population $\delta/\sigma = 2/e = .7358$, so similar results are obtainable.

For various geometric models one has : (a) rectangular, $\delta/\sigma = \sqrt{3/4} = .87$, (b) triangular, $\delta/\sigma = \sqrt{2/3} = .82$, (c) right triangular, $\delta/\sigma = 16\sqrt{2/27} = .84$, and (d) V-shaped, $\delta/\sigma = \sqrt{8/9} = .94$. All values of δ/σ are higher than for a normal population so that the inequalities based on δ would be better than those based on σ only in more limited cases.

For Pareto-type distributions, however, the inequalities based on δ will very often be sharper. If the Pareto coefficient, α , satisfies $1.0 < \alpha \leq 2.0$, δ exists but σ does not (a value of about $\alpha = 1.5$ is often observed). And even if $\alpha > 2$, δ/σ is likely to be small for

$$\delta/\sigma = 2 \left(\frac{\alpha-1}{\alpha} \right)^{\alpha-1} \sqrt{\frac{\alpha-2}{\alpha}}.$$

When $\alpha = 3$, the ratio is .51 while if $\alpha = 5$, it is .63. This model therefore provides an illustration of the type of situation in which inequalities based on δ are better over a wide range because the value of σ is so much more affected by extreme deviations from the mean.

4. APPLICATIONS

Applications of Tchebycheff-type inequalities are well known and need not be discussed here in detail. It may be noted, however, that they are often useful in analyzing secondary data where estimates of only a few descriptive parameters have been made available. The results given here, therefore, may have some significance for those who report such data when it is desirable to give a measure of variability. Unless some population model can be assumed, the choice of which measure of variability to report is often arbitrary. The results given here suggest that if it is known that the population is one with highly extreme deviations from the mean the value of δ might well be somewhat more meaningful, in at least one way, to users of the data than would the value of σ .

REFERENCES

- CHAKRAVARTI, M. C. (1947-48): On the ratio of the mean deviation to the standard deviation. *Bull. Cal. Stat. Ass.*, **1**, 187-190.
- GODWIN, H. J. (1955): On generalizations of Tchebycheff's inequality. *J. Amer. Stat. Ass.*, **50**, 923-945.
- HOEFFDING, W. (1952): The large-sample power of tests based on permutations of observations. *Ann. Math. Stat.*, **23**, 169-192.
- SELBERG, H. L. (1940): Zwei ungleichungen zur ergänzung des Tchebycheffschen lemmas. *Skandinavisk Aktuarietidskrift*, **23**, 121-125.

Paper received: August, 1961.

A NOTE ON FUNCTIONAL MINIMALITY OF SUFFICIENT STATISTICS*

By EDWARD W. BARANKIN

University of California, Berkeley

SUMMARY. Theorem 4.3 in the article of Barankin and Katz (1959) states that the minimal dimensional sufficient statistics T^* and T^+ , there constructed, are locally functionally minimal almost everywhere in R ($\equiv$ the set of regular points), whereas in fact, the correct assertion is that they are locally *almost everywhere* functionally minimal almost everywhere in R . This latter, valid statement is proved here. Also, we give corrected forms of certain results that were derived, in our work on sufficient parameters, from the misstated Theorem 4.3.

1. INTRODUCTION

The purpose of this note is to correct an erroneous assertion made in the article "Sufficient Statistics of Minimal Dimension" by Barankin and Katz (1959). Theorem 4.3, in that article states that the sufficient statistics T^* and T^+ , which were constructed earlier, are locally functionally minimal at every point of an open set $A \subseteq R$ ($\equiv$ the set of regular points of Ω) with $R-A$ of Lebesgue measure 0. This is explained to mean that for each point $x \in A$ there is a neighbourhood N of x such that if T is any sufficient statistic then T^* and T^+ are functions of T in N . Now, with this meaning, the affirmation of local functional minimality is not strictly correct in general. It is true in special cases; for example, by virtue of Lemma 2.3 in B. K. (we shall henceforth refer to the above quoted article as B. K.) it is true if T is Euclidean and regular in A . (The reason for this will become apparent with the ensuing discussion.) But in general, if T is just any sufficient statistic according to the characterization given by Lemma 1.1 in B. K., the correct assertion is that T^* and T^+ are locally *almost everywhere* functionally minimal everywhere in A . That is, for each $x \in A$ there is a neighbourhood N of x such that if T is any sufficient statistic then *there is a subset C_N of N , of Lebesgue measure 0, such that $T^*(x') = T^*(x'')$ and $T^+(x') = T^+(x'')$ whenever x' and x'' are points of $N - C_N$ with $T(x') = T(x'')$.*

The proof of Theorem 4.3 in B. K. was reduced to the following step: for a certain collection of indices, $j_1, j_2, \dots, j_r$, and a corresponding collection of points, $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(r)}$, of Θ , to show that each of the functions

$$\left(\frac{\partial \log p}{\partial \theta_{j_k}} \right)_{x, \theta^{(k)}} \quad \dots \quad (1.1)$$

is a function of an arbitrary, given sufficient statistic T . To establish this we proceeded from the characterizing equation for the sufficient statistic T :

$$p(x, \theta) = f(T(x), \theta)g(x). \quad \dots \quad (1.2)$$

*This paper was prepared with the partial support of the Office of Naval Research (Nonr 222-43). This paper in whole or in part may be reproduced for any purpose of the United States Government.

taking the logarithms, differentiating with respect to θ_{j_k} , and then evaluating at $\theta^{(k)}$. This gives the result

$$\left(\frac{\partial \log p}{\partial \theta_{j_k}} \right)_{x, \theta^{(k)}} = \left(\frac{\partial \log f(T(x), \cdot)}{\partial \theta_{j_k}} \right)_{\theta^{(k)}}, \quad \dots \quad (1.3)$$

from which it then appears as evident that (1.1) is strictly a function of T . But this procedure is not generally valid. And this is because (1.2) is not given to hold for all x and θ . From Lemma 1.1 in B. K. we see that what we may assert, in general, regarding (1.2) is that for each $\theta \in \Theta$ it holds for almost all (Lebesgue) $x \in \Omega$. Under these circumstances we cannot draw the conclusion that f has the θ -differentiability properties that are needed to obtain (1.3). Thus, the result of strict functionality cannot be asserted.

If T is Euclidean and regular at a point x^0 , then Lemma 2.3 in B. K. provides us with functions f and g in (1.2) which are suitably differentiable and such that (1.2) holds for all x in some neighbourhood of x^0 and all $\theta \in \Theta$. In this case, the above argumentation is valid for all x in a neighbourhood of x^0 , and one obtains the correct result that (1.1) is strictly a function of T locally about x^0 . This is, of course, a very special situation.

In Section 2 below we shall state and prove the valid form of Theorem 4.3 of B. K.

There has so far appeared in the literature only one application of the incorrect Theorem 4.3; this is in the article of Barankin (1960). There, in Theorem 5.5 and Corollary 5.5.1, we have obtained consequences regarding the identifiability of parameters. In Section 3 below we shall give corrected statements of these parametric results. Once again it will be a matter merely of substituting almost everywhere functionality for everywhere functionality.

We ask the reader's indulgence in having before him, as he reads this note, the two papers referred to above (see the References at the end of the note). Indeed, if we were to attempt to make this note self-contained, we should have to devote several times the present number of pages to purely preliminary discussion.

2. CORRECTION OF THEOREM 4.3., IN B. K. (1959)

We wish to prove now the following :

Theorem 4.3 (B. K., 1959, corrected form) : *The $\mathcal{P}$ -sufficient statistics T^* and T^+ , of the preceding two theorems, are locally almost everywhere (Lebesgue) functionally minimal at almost all (Lebesgue) points of R . That is, for each point x in a set $A \subseteq R$, with $R - A$ of measure 0, there is a neighbourhood N of x such that if T is any sufficient statistic then there is a subset C_N of N of measure 0, such that $T^*(x') = T^*(x'')$ and $T^+(x') = T^+(x'')$ whenever x' and x'' are points of $N - C_N$ with $T(x') = T(x'')$.*

Proof : It is readily seen that the argument adduced in B. K. to derive the stated assertion for T^+ from that for T^* is applicable in the present situation. That is, in a word, T^+ is strictly a function of T^* on each K_{si}^0 , and therefore if T^* is almost everywhere a function of T in K_{si}^0 then it follows that the same is true of T^+ .

Let us then consider T^* . The set A is taken to be the disjoint union of open (n -dimensional) intervals, $\bigcup_{s,i} K_{si}^0$. For a point x in one of the K_{si}^0 , we take this K_{si}^0 to be the neighbourhood N spoken of in the theorem. On the cell K_{si} the statistic T^* is, for some r ,

A NOTE ON FUNCTIONAL MINIMALITY OF SUFFICIENT STATISTICS

an r -dimensional vector function whose components are of the form (1.1) for some j_k and $\theta^{(k)}$, $k = 1, 2, \dots, r$. Since there are only countably many of these component functions, it clearly suffices to show that, if T is any sufficient statistic, then each of the functions is almost everywhere a function of T in K_{si} . And this will follow *a fortiori* if we show that any function of the form (1.1) is almost everywhere a function of T in Ω . It is this that we now proceed to establish.

Let j_0 be any particular integer between 1 and v , inclusive, and θ^0 be any particular point of Θ ; and define

$$\eta(x) = \left(\frac{\partial \log p}{\partial \theta_{j_0}} \right)_{x, \theta^0} \quad \dots \quad (2.1)$$

Let T be a sufficient statistic; thus, there are functions f and g such that, for each $\theta \in \Theta$, (1.2) holds for almost all $x \in \Omega$. Specifically, for a given θ , let C_θ denote the set of points $x \in \Omega$ for which (1.2) does not hold. C_θ has, then, Lebesgue measure 0, for each θ .

If t is any real number, let the point $\theta(t)$ be defined as follows :

$$\theta(t) = (\theta_1^0, \theta_2^0, \dots, \theta_{j_0-1}^0, \theta_{j_0}^0 + t, \theta_{j_0+1}^0, \dots, \theta_v^0). \quad \dots \quad (2.2)$$

Thus, in particular, $\theta(0) = \theta^0$. Since Θ is an open set there is a positive number t_0 such that $\theta(t) \in \Theta$ for all t in the interval $(0, t_0)$. Let $t_1 > t_2 > \dots$ be a monotone decreasing sequence of numbers in $(0, t_0)$ converging to 0. Let $C = C_{\theta^0} \cup \bigcup_{m=1}^{\infty} C_{\theta(t_m)}$. Clearly C has Lebesgue measure 0. And furthermore, if $x \in \Omega - C$ then (1.2) holds for $\theta = \theta^0$ and $\theta = \theta(t_m)$ for all $m = 1, 2, \dots$

According to this last property of C we obtain the following from (1.2) :

$$\frac{\log p(x, \theta(t_m)) - \log p(x, \theta^0)}{t_m} = \frac{\log f(T(x), \theta(t_m)) - \log f(T(x), \theta^0)}{t_m}, \quad \dots \quad (2.3)$$

$$m = 1, 2, \dots; x \in \Omega - C.$$

As $m \rightarrow \infty$, the left-hand side of (2.3) approaches $\eta(x)$. Therefore, the limit of the quantities on the right-hand side exists also, and is likewise $\eta(x)$. But we see that the quantities on the right form identical sequences for two points, x' and x'' , in $\Omega - C$ such that $T(x') = T(x'')$. It therefore follows that η is a function of T on $\Omega - C$. Since C is of measure 0 we have that η is almost everywhere a function of T in Ω , and this completes the proof of the theorem.

3. CORRECTION OF THEOREMS 2.5 AND 5.5 AND COROLLARY 5.5.1 IN BARANKIN (1960)

In the article B. (1960) (we shall henceforward refer in this way to the second item in the References at the end of this note) we showed that sufficient parameters of our family of distributions, $\mathcal{P}$, may be viewed as sufficient statistics of a related posterior family, and conversely. (However, regarding the converse, see the discussion on p. 108 et seq. of B. (1960)). This remarkable fact enabled us to carry all of the article B. K. (1959) over to questions concerning sufficient parameters. And in particular, assertions about functional minimality of sufficient statistics yield assertions about identifiability of parameters.

The statement of Theorem 4.3 for T^* was quoted in B. (1960) as Theorem 2.5. We first give here the proper restatement of that theorem.

Theorem 2.5 (B. 1960, corrected form) : *The sufficient statistic T^* of the preceding theorem is locally almost everywhere (Lebesgue) functionally minimal at almost all (Lebesgue) points of R . That is, there is an open set $A \subseteq R$, with $R-A$ of Lebesgue measure 0, such that for each $x \in A$ there is a neighbourhood N of x with the property that if T is any sufficient statistic for $\mathcal{P}$ in N then there is a subset C_N of N , of measure 0, such that $T^*(x') = T^*(x'')$ whenever x' and x'' are points of $N-C_N$ with $T(x') = T(x'')$.*

It will be noticed that this statement does not refer to a sufficient statistic, T , for $\mathcal{P}$, but to a sufficient statistic for $\mathcal{P}$ in N . We have defined this slightly more general notion in Definition 2.1 of B. (1960). The reader will have no difficulty in seeing that the present theorem is implied by the theorem of Section 2 above; it is necessary only to observe that if T is sufficient for $\mathcal{P}$ in N then the function which is equal to T in N and equal to the identity function in $\Omega-N$ is a sufficient statistic for $\mathcal{P}$.

In B. (1960) we gave three definitions bearing on the notion of identifiability. It will be necessary now to introduce also the notion of almost everywhere identifiability. Let us give formal definitions as follows (cf. Definitions 5.2, 5.3 and 5.4 in B. (1960)) :

Definition : A sufficient parameter, U , for $\mathcal{P}$ is said to be almost everywhere (Lebesgue) identifiable if there is a subset D of Θ of (Lebesgue) measure 0, such that

$$\mu_\theta = \mu_{\theta'} \quad \text{implies} \quad U(\theta) = U(\theta') \quad \dots \quad (3.1)$$

for all θ and θ' in $\Theta-D$.

A parameter U of $\mathcal{P}$ which is sufficient for $\mathcal{P}$ in F (a measurable subset of Θ), is said to be almost everywhere identifiable in F if there is a subset D of F , of measure 0, such that (3.1) holds for all θ and θ' in $F-D$.

A parameter U of $\mathcal{P}$, which is sufficient for $\mathcal{P}$ about θ^0 , is said to be almost everywhere identifiable about θ^0 if it is almost everywhere identifiable in some neighbourhood of θ^0 .

The reader is referred to B. (1960) for all attendant notions, such as sufficiency of a parameter in a given set and sufficiency of a parameter about a given point.

We are now ready to state the corrected forms of Theorem 5.5 and Corollary 5.5.1. And it will not be necessary to supply any revised detail of proof. For, the argumentation given in B. (1960) remains fully valid when almost everywhere identifiability is substituted for identifiability and the above corrected form of Theorem 2.5 is substituted for the invalid one.

Theorem 5.5 (B., 1960, corrected form) : *The sufficient parameter U^* of the preceding theorem is locally almost everywhere (Lebesgue) functionally minimal at almost all (Lebesgue) points of R_Θ . That is, there is an open set $B \subseteq R_\Theta$, with $R_\Theta-B$ of Lebesgue measure 0, such that for each $\theta \in B$ there is a neighbourhood N of θ with the property that if U is any sufficient parameter for $\mathcal{P}$ in N then there is a subset D_N of N , of measure 0, such that $U^*(\theta') = U^*(\theta'')$ whenever θ' and θ'' are points of $N-D_N$ with $U(\theta') = U(\theta'')$.*

Corollary 5.5.1 (B., 1960, corrected form) : *The sufficient parameter U^* is almost everywhere (Lebesgue) identifiable about each point $\theta \in B$.*

REFERENCES

- BARANKIN, E. W. and KATZ, M. (1959) : Sufficient statistics of minimal dimension. *Sankhyā*, **21**, 217-246.
 BARANKIN, E. W. (1960) : Sufficient parameters : solution of the minimal dimensionality problem. *Ann. Inst. Stat. Math.*, Tokyo, **XII**, 91-118.

Paper received : July, 1961.

MINIMAX SOLUTIONS TO TRICHOTOMIES

By L. L. LASMAN

Florida State University

SUMMARY. When a three-valued hypothesis is to be tested and actions are available for accepting one from the three, it is shown that under rather general conditions, the minimax solution to this problem depends upon a pair of values x' and x'' with $x' < x''$. These are obtained by equating the risks under possible alternatives. The decision then is to choose the smallest hypothesized parameter if $\Sigma x \leq x'$, choose the largest if $\Sigma x > x''$ and choose the middle value otherwise.

1. INTRODUCTION

The explicit method of finding Bayes' and minimax solutions in the problem of testing dichotomies is quite simple and well known for a very general class of probability distributions and is set out by Blackwell and Girshick (1954). An article by Karlin and Rubins (1956) contains a general theory of the class of Bayes' solutions, and this paper presents an explicit method of obtaining minimax solutions to trichotomies for a more specific family of distributions.

The distribution to be considered is one of the exponential type

$$f(x, \theta) = b(\theta)e^{\theta x}g(x), \quad \dots (1.1)$$

where $b(\theta)$ is a positive function of a real-valued parameter θ , and $g(x)$ is a positive function of the real variable x . The normal distribution with known variance, and the binomial and Poisson distributions are of this type.

We are concerned with testing a three-valued hypothesis, that is, with the problem of accepting one of the three following hypotheses:

$$\left. \begin{array}{l} H_1: \theta = \theta_1 \\ H_2: \theta = \theta_2 \\ H_3: \theta = \theta_3 \end{array} \right\}$$

where we may assume $\theta_1 < \theta_2 < \theta_3$.

Let $x_1, \dots, x_n$ denote the value of n independent observations on x ; let a_i denote the action which accepts H_i and let $\phi(a_i|x)$ denote the probability of taking action a_i , given the sample point $x = (x_1, \dots, x_n)$. We require that for each x ,

$$0 \leq \phi(a_i|x) \leq 1 \quad \text{and} \quad \Sigma \phi(a_i|x) = 1.$$

If in particular $\phi(a_i|x) = 1$, for only one i , then we have the customary non-randomized test procedure.

Let w_{ij} represent the loss incurred from taking action a_i when the hypothesis H_j is actually true. We will assume that the w_{ij} are non-negative, that there is no loss from choosing the most appropriate action corresponding to the actual

parameter value, and that there cannot be a smaller penalty if one accepts a hypothesis farther from the correct one. This defines a loss matrix as:

$a \backslash \theta$	θ_1	θ_2	θ_3
a_1	0	w_{12}	w_{13}
a_2	w_{21}	0	w_{23}
a_3	w_{31}	w_{32}	0

where $w_{13} \geq w_{23}$ and $w_{31} \geq w_{21}$.

Let ξ_i , $i = 1, 2, 3$, be the *a priori* probability that θ_i is the true value of θ , where $\xi_i \geq 0$ and $\sum \xi_i = 1$. Denote the vector (ξ_1, ξ_2, ξ_3) by ξ . We shall define the risk corresponding to nature's choice of an *a priori* distribution ξ and the statistician's choice of a test procedure ϕ as

$$\begin{aligned} R(\xi, \phi) &= \sum_{j=1}^3 \xi_j \int \dots \int \sum w_{ij} \phi(a_i | x) f(x, \theta_j) dx \\ &= \sum_{j=1}^3 \xi_j R(\theta_j, \phi) = \int \dots \int \sum_{i=1}^3 \phi(a_i | x) \rho(a_i | x) g(x) dx \end{aligned} \quad (1.2)$$

where the region of integration is the sample space. $R(\theta_i, \phi)$ is the risk from choosing ϕ when θ_i is the value of the parameter and $\rho(a_i | x)$ the risk in taking action i after observing x . In particular

$$\rho(a_i | x) = \sum_{j=1}^3 w_{ij} \xi_j e^{\theta_j \Sigma x} b(\theta_j). \quad \dots (1.3)$$

Then it is known that there exists a pair (x', x'') with $x' < x''$ such that the Bayes' solution, one which minimizes $R(\xi, \phi)$, is choosing a_1 when $\Sigma x \leq x'$, a_2 when $x' < \Sigma x \leq x''$, and a_3 when $\Sigma x > x''$. For $\Sigma x = x'$ or $\Sigma x = x''$, randomization may be required in the discrete cases. The numbers x' and x'' have been obtained such that $\rho(a_1 | x) = \rho(a_2 | x)$ for $\Sigma x = x'$ and $\rho(a_2 | x) = \rho(a_3 | x)$ for $\Sigma x = x''$.

2. THE MINIMAX SOLUTION

By a minimax test ϕ^* , we mean that

$$\sup_{\xi} R(\xi, \phi^*) \leq \sup_{\xi} R(\xi, \phi) \text{ for all } \phi.$$

For a test of the form of the Bayes' solution just given, the functions $R(\theta_i, \phi)$ can be written as

$$R(\theta_1, \phi) = w_{21} P_{\theta_1}(x' < \Sigma x \leq x'') + w_{31} P_{\theta_1}(\Sigma x > x'') \quad \dots (2.1)$$

$$R(\theta_2, \phi) = w_{12} P_{\theta_2}(\Sigma x \leq x') + w_{32} P_{\theta_2}(\Sigma x > x'') \quad \dots (2.2)$$

$$R(\theta_3, \phi) = w_{13} P_{\theta_3}(\Sigma x \leq x') + w_{23} P_{\theta_3}(x' < \Sigma x \leq x''). \quad \dots (2.3)$$

MINIMAX SOLUTIONS TO TRICHOTOMIES

It will now be shown that if a pair (x', x'') with $x' < x''$ exists so that equality holds among the three risks, then the Bayes' test above is minimax. Call such a test ϕ^* . Then we can observe that for this ϕ^*

$$R(\xi, \theta) = \sum \xi_i R(\theta_i, \phi^*) = V, \quad \sum \xi_i = V, \quad \dots \quad (2.4)$$

where V is the common risk to the statistician regardless of which hypothesis is true.

It remains to be shown that there is an *a priori* distribution ξ^* such that

$$R(\xi^*, \phi) \geq R(\xi^*, \phi^*) = V \quad \dots \quad (2.5)$$

for we have found that $V = \sup_{\xi} R(\xi, \phi^*)$.

Let $z = e^{\Sigma x}$, $z' = e^{\Sigma x'}$ and $z'' = e^{\Sigma x''}$. It must be shown that for the choice of x' and x'' such that equality holds among equations (2.1), (2.2), and (2.3), there corresponds a vector (ξ_1, ξ_2, ξ_3) of *a priori* probabilities satisfying $\rho(a_1|x) = \rho(a_2|x)$ for $z = z'$ and $\rho(a_2|x) = \rho(a_3|x)$ for $z = z''$. But this is equivalent to showing that there exists a positive point in the ξ_1', ξ_2' -plane corresponding to the choice of (x_1', x'') , where $\xi_1' = \xi_1/\xi_3$ and $\xi_2' = \xi_2/\xi_3$.

Dividing each $\rho(a_i|x)$ by $\xi_3 z^{\theta_3}$, setting $\rho(a_1|x) = \rho(a_2|x)$ for $z = z'$ and $\rho(a_2|x) = \rho(a_3|x)$ for $z = z''$, a solution for ξ_1' and ξ_2' can be obtained. These, in determinantal form, are

$$\xi_1' = \frac{\begin{pmatrix} b(\theta_3)(w_{13}-w_{23}) & -w_{12}b(\theta_2)z^{(\theta_2-\theta_3)'} \\ -b(\theta_3)w_{23} & -w_{32}b(\theta_2)z^{(\theta_2-\theta_3)''} \end{pmatrix}}{D} = \frac{D_1}{D}$$

$$\xi_2' = \frac{\begin{pmatrix} w_{21}b(\theta_1)z^{(\theta_1-\theta_3)'} & b(\theta_3)(w_{13}-w_{23}) \\ b(\theta_1)(w_{21}-w_{31})z^{(\theta_1-\theta_3)''} & -b(\theta_3)w_{23} \end{pmatrix}}{D} = \frac{D_2}{D}$$

where

$$D = \begin{pmatrix} w_{21}b(\theta_1)z^{(\theta_1-\theta_3)'} & -w_{12}b(\theta_2)z^{(\theta_2-\theta_3)'} \\ b(\theta_1)(w_{21}-w_{31})z^{(\theta_1-\theta_3)''} & -w_{32}b(\theta_2)z^{(\theta_2-\theta_3)'} \end{pmatrix}$$

Both D and D_1 are easily shown to be negative. In order for ξ_2' to be positive, D_2 must be negative. This is so if

$$w_{21}, w_{23}z^{(\theta_1-\theta_3)'} > (w_{31}-w_{21})(w_{13}-w_{23})z^{(\theta_1-\theta_3)''}$$

and this inequality is implied by the hypothesis that a pair (x', x'') exists such that $R(\theta_1, \phi) = R(\theta_2, \phi) = R(\theta_3, \phi)$.

Now $\xi_1' > 0$ and $\xi_2' > 0$. This defines the triple (ξ_1, ξ_2, ξ_3) , and it is easily seen that it has the properties of a probability distribution.

Suppose that this distribution is ξ^* . Then by construction of ξ^* , the Bayes' solution against it is of the form specified in Section 1. This is the minimax solution since the risk to the statistician if he uses ϕ^* is at most V , and if nature uses the distribution ξ^* , then the smallest risk to the statistician is V corresponding to the Bayes' risk against this distribution.

3. AN EXAMPLE

As a simple example consider the normal distribution with variance known to be unity and the following set of hypotheses:

$$H_1: \mu = -1$$

$$H_2: \mu = 0$$

$$H_3: \mu = 1$$

Assume a simple loss matrix as

$a \backslash \mu$	-1	0	1
a_1	0	1	2
a_2	1	0	1
a_3	2	1	0

The risks $R(\theta_i, \phi)$ when a single observation is taken then become

$$R(\theta_1, \phi) = \int_{x'}^{x''} f(x, -1)dx + 2 \int_{x''}^{\infty} f(x, -1)dx$$

$$R(\theta_2, \phi) = \int_{-\infty}^{x'} f(x, 0)dx + \int_{x''}^{\infty} f(x, 0)dx$$

$$R(\theta_3, \phi) = 2 \int_{-\infty}^{x'} f(x, 1)dx + \int_{x'}^{x''} f(x, 1)dx.$$

Equating $R(\theta_1, \phi)$ to $R(\theta_3, \phi)$ defines points on opposite sides of and equidistant from 0. Setting these equal to $R(\theta_2, \theta)$ gives unique values to these numbers. Using tables of the normal distribution it is readily found that $x' = -.763$ and $x'' = .763$. Thus the minimax solution is to take the observation and if it is less than $-.763$ accept the mean as -1 , if greater than $.763$, choose $\mu = 1$ and $\mu = 0$ otherwise.

A practical application of the method using a discrete distribution might be in the area of acceptance sampling where a purchaser has three alternative actions to choose from : (1) accept the lot, (2) accept the lot at a discount when it may be composed of "seconds", (3) reject the lot. The lot may be good, poor, or bad depending on the value of the proportion defective, $p_1 < p_2 < p_3$. In such a case, randomization may be necessary.

REFERENCES

- BLACKWELL, D. and GIRSHICK, M. A. (1954): *Theory of Games and Statistical Decisions*, John Wiley and Sons, New York.
- KARLIN, S. and RUBIN, H. (1956): The theory of decision procedures for distributions with monotone likelihood ratio. *Ann. Math. Stat.*, 27, 272-299.

Paper received : July, 1960.

USE OF 'ORDER-STATISTIC' IN SAMPLING WITHOUT REPLACEMENT

By P. K. PATHAK

Indian Statistical Institute

SUMMARY. In sampling without replacement from a finite population, the order in which the units are selected, is immaterial for the purpose of estimation. This point was noted by Basu (1958) and Murthy (1957). Basu showed that the 'order-statistic' (sample units arranged in ascending order of their unit indices) forms a sufficient statistic, and, therefore, any estimator which is not a function of the order-statistic, can be uniformly improved by the use of Rao-Blackwell theorem. In this paper, certain results obtained by Murthy are shown to be immediate consequences of the above observation.

It is shown that sampling with different probabilities with replacement, until we get a specified number of distinct units, is equivalent in some sense to sampling with different probabilities without replacement. Some other related problems are also considered here.

1. INTRODUCTION

Let $Y_1, Y_2, \dots, Y_N$ be the Y -characteristic of the N population units under study. Let P_j be the probability associated with the j -th population unit ($j = 1, \dots, N$). For simplicity, we shall always refer population units by capital letters and sample units by small letters, e.g., y_i and p_i denote the variate value and the probability of selection associated with the i -th sample unit respectively. In this paper, we shall throughout follow the notations used by Basu (1958).

2. SAMPLING WITHOUT REPLACEMENT

In sampling without replacement from the above population a particular sample may be recorded as

$$s = (x_1, x_2, \dots, x_n),$$

where $x_i = (y_i, p_i, u_i)$ ($i = 1, 2, \dots, n$), and n is the sample size.

The probability of drawing such a particular sample is given by

$$P(s) = p_1 \cdot \frac{p_2}{(1-p_1)} \cdot \frac{p_3}{(1-p_1-p_2)} \cdots \frac{p_n}{(1-p_1-\cdots-p_{n-1})} \cdots \quad (2.1)$$

If we record the 'order-statistic' by

$$T = (x_{(1)}, \dots, x_{(n)}),$$

where $x_{(i)} = (y_{(i)}, p_{(i)}, u_{(i)})$ is the i -th order-statistic ($i = 1, \dots, n$), we have

$$P(T) = \sum_{s \in T} \frac{p_1 p_2 \dots p_n}{(1-p_1)(1-p_1-p_2) \dots (1-p_1-p_2-\dots-p_{n-1})}, \quad \dots \quad (2.2)$$

where the summation is taken over all possible samples giving rise to the 'order-statistic' T .

It has been shown by Basu (1958) that T is a sufficient statistic. Thus, if $g(s)$ is some estimator depending only on s , by Rao-Blackwell theorem, a uniformly better estimator than $g(s)$ is given by $E(g(s)|T)$. For any convex loss function, the risk associated with $E(g(s)|T)$ is smaller than the risk associated with $g(s)$.

3. SAMPLING WITH REPLACEMENT : NUMBER OF DISTINCT UNITS FIXED IN ADVANCE

In this case, units are drawn with unequal probabilities and with replacement until we get a specified number ' n ' of distinct units. If r denotes the number of draws in a particular case, the sample s may be recorded as

$$s = (x_1, \dots, x_r).$$

If we define the 'order-statistic' T by

$$T = [x_{(1)}, x_{(2)}, \dots, x_{(n)}],$$

where $x_{(i)}$ is the i -th 'order-statistic' ($i = 1, 2, \dots, n$), it is not difficult to show that

$$P(T) = \sum_{r=2}^{\infty} \left[\sum_{i=1}^n p_{(i)} \{ \sum^{(i)} (p_{(1)} + \dots + p_{(n)})^{r-1} - \sum^{(i)} (p_{(1)} + \dots + p_{(n-1)})^{r-1} \right. \\ \left. + \dots + (-)^{n-2} \sum^{(i)} p_{(1)}^{r-1} \} \right], \quad \dots \quad (3.1)$$

where $\sum^{(i)}$ denotes the summation over all possible combinations out of $p_{(1)}, \dots, p_{(i-1)}, p_{(i+1)}, \dots, p_{(n)}$, and the term inside the square brackets denotes the probability of getting T in r draws. Assuming without any loss of generality that $p_{(1)} + \dots + p_{(n)} < 1$, we get on summing (3.1) over r

$$P(T) = \sum_{i=1}^n p_{(i)} \left[\sum^{(i)} \frac{p_{(1)} + \dots + p_{(n)}}{1 - p_{(1)} - \dots - p_{(n)}} - \sum^{(i)} \frac{p_{(1)} + \dots + p_{(n-1)}}{1 - p_{(1)} - \dots - p_{(n-1)}} \right. \\ \left. + (-)^{n-2} \sum^{(i)} \frac{p_{(1)}}{1 - p_{(1)}} \right]. \quad \dots \quad (3.2)$$

USE OF ORDER-STATISTIC IN SAMPLING WITHOUT REPLACEMENT

It can be proved by induction over n that (3.2) and (2.2) are equal.

Thus, if we rely only on the 'order-statistic' T , the two methods of sampling are essentially the same.

4. IMPROVING DES RAJ'S ESTIMATORS

Des Raj (1956), in sampling without replacement, gave the following set of uncorrelated estimators of $Y = \sum_{i=1}^X Y_i$.

$$t_1(s) = y_1/p_1;$$

$$t_2(s) = y_1 + \frac{y_2}{p_2} (1 - p_1);$$

$$\dots \quad \dots \quad \dots \quad \dots \quad (4.1)$$

$$t_i(s) = y_1 + y_2 + \dots + y_{i-1} + \frac{y_i}{p_i} (1 - p_1 - p_2 - \dots - p_{i-1});$$

• • • • •

$$t_n(s) = y_1 + y_2 + \dots + y_{n-1} + \frac{y_n}{p_n}(1 - p_1 - p_2 - \dots - p_{n-1}).$$

Theorem 1 : *For any convex loss function a uniformly better estimator than*

$$t(s) = \sum_{i=1}^n c_i t_i(s) \quad \left(\sum_{i=1}^n c_i = 1 \right)$$

is given by

$$E[t(s)|T] = \sum_{i=1}^n y_{(i)} \frac{P(T/(i))}{P(T)}, \quad \dots \quad (4.2)$$

where $P(T|(i))$ is the conditional probability of getting the 'order-statistic' T given that i -th order unit was drawn first.

Proof : Since

$$\begin{aligned} & \frac{P[x_i = x_{(j)}, x_{i+1} = x_{(k)} \mid x_1, x_2, \dots, x_{i-1}]}{P[x_i = x_{(k)}, x_{i+1} = x_{(j)} \mid x_1, x_2, \dots, x_{i-1}]} \\ &= \frac{1 - p_1 - p_2 - \dots - p_{i-1} - p_{(k)}}{1 - p_1 - p_2 - \dots - p_{i-1} - p_{(j)}}, \dots \quad \dots \quad (4.3) \end{aligned}$$

it follows that

$$E[t_{i+1}(s) - t_i(s) | T] = E\left[\frac{y_{i+1}}{p_{i+1}}(1 - p_1 - p_2 - \dots - p_i) - \frac{y_i}{p_i}(1 - p_1 - p_2 - \dots - p_i) | T\right] = 0.$$

Therefore,

$$E \left[\sum_{i=1}^n c_i t_i(s) | T \right] = E[t_1(s) | T] = \sum_{i=1}^n y_{(i)} \frac{P(T|(i))}{P(T)}.$$

Corollary 1 : When $n = 2$, we have

$$E \left[\sum_{i=1}^2 c_i t_i(s) | T \right] = \frac{1}{(2-p_1-p_2)} \left[(1-p_2) \frac{y_1}{p_1} + (1-p_1) \frac{y_2}{p_2} \right].$$

Corollary 2 : In simple random sampling without replacement

$$t_i(s) = y_1 + \dots + y_{i-1} + (N-i+1) y_i, \quad i = 1, \dots, n;$$

and

$$E[t_i(s) | T] = E[t_1(s) | T] = \frac{N}{n} \sum_{i=1}^n y_i.$$

Theorem 2 : A uniformly better estimator than

$$g(s) = \sum_{i \neq j=1}^n c_{ij} t_j(s) t_j(s)$$

($\sum_{i \neq j=1}^n c_{ij} = 1$) of Y^2 , is given by

$$g(T) = E[g(s) | T] = \frac{\sum_{i=1}^n y_{(i)}^2 P(T|(i)) + \sum_{i \neq i'=1}^n y_{(i)} y_{(i')} P(T|(i), (i'))}{P(T)}. \quad \dots (4.4)^*$$

Proof : Using (4.3), it can be seen that

$$E[t_1(s) t_{j+1}(s) - t_1(s) t_j(s) | T] = 0, \quad (j = 2, \dots, n)$$

and

$$E[t_{i+1}(s) t_j(s) - t_i(s) t_j(s) | T] = 0, \quad (i = 1, 2, \dots, j-2)$$

and hence

$$E[g(s) | T] = E[t_1(s) t_2(s) | T]$$

$$= \frac{\sum_{i=1}^n y_{(i)}^2 P(T|(i)) + \sum_{i \neq i'=1}^n y_{(i)} y_{(i')} P(T|(i), (i'))}{P(T)}$$

* *Remark :* The estimators (4.2) and (4.4) can also be obtained by improving the usual estimators of Y and Y^2 under the sampling scheme discussed in Section 3.

USE OF ORDER-STATISTIC IN SAMPLING WITHOUT REPLACEMENT

Corollary 3 : When $n = 2$, we have

$$E[t_1(s) t_2(s) | T] = \frac{1}{(2-p_{(1)}-p_{(2)})} \left[(1-p_{(2)}) \frac{y_{(1)}^2}{p_{(1)}} + (1-p_{(1)}) \frac{y_{(2)}^2}{p_{(2)}} + 2(1-p_{(1)})(1-p_{(2)}) \frac{y_{(1)}}{p_{(1)}} \frac{y_{(2)}}{p_{(2)}} \right].$$

Corollary 4: In simple random sampling (without replacement)

$$E[t_i(s) t_j(s) | T] = \frac{N}{n} \sum_{i=1}^n y_{(i)}^2 + \frac{N(N-1)}{n(n-1)} \sum_{i \neq j=1}^n y_{(i)} y_{(j)}.$$

The estimator (4.4) is used to derive unbiased variance estimator of

$$\sum_{i=1}^n y_{(i)} \frac{P(T | (i))}{P(T)}.$$

5. IMPROVING DAS' ESTIMATORS

The set of estimators of Y given by Das (1951) is as follows :

$$\begin{aligned}
 u_1(s) &= \frac{y_1}{p_1} ; \\
 u_2(s) &= \frac{y_2}{p_2} \cdot \frac{(1-p_1)}{p_1} \cdot \frac{1}{(N-1)} ; \\
 &\dots \qquad \dots \qquad \dots \qquad \dots \\
 u_r(s) &= \frac{y_r}{p_r} \cdot \frac{(1-p_1-p_2 \dots -p_{r-1})}{p_{r-1}} \cdot \dots \cdot \frac{(1-p_1-p_2)}{p_2} \cdot \frac{1-p_1}{p_1} \cdot \\
 &\qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \frac{1}{(N-1)(N-2) \dots (N-r+1)} ; \\
 &\dots \qquad \dots \qquad \dots \qquad \dots \qquad \dots \\
 u_n(s) &= \frac{y_n}{p_n} \cdot \frac{(1-p_1-p_2 \dots -p_{n-1})}{p_{n-1}} \cdot \dots \cdot \frac{(1-p_1-p_2)}{p_2} \cdot \frac{(1-p_1)}{p_1} \cdot \\
 &\qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \frac{1}{(N-1)(N-2) \dots (N-n+1)} \cdot \dots \quad (5.1)
 \end{aligned}$$

A uniformly better estimator than $u_r(s)$ is, therefore, given by

$$u_r(T) = E[u_r(s) | T]$$

$$= \sum_{i=1}^n y_{(i)} \frac{\sum' \frac{1}{(N-1)(N-2)\dots(N-r+1)} P(T | x_1, x_2, \dots, x_{r-1}, x_r = x_{(i)})}{P(T)}$$

$$r = 1, \dots, n \quad (5.2)$$

where the summation Σ' is taken over all possible $x_1, \dots, x_{r-1}$.

It is easy to see that the estimators $u_r(T)$ ($r = 1, \dots, n$) are identical if and only if the sample is drawn by simple random sampling (without replacement). In this case (5.2) is same as (4.2).

This shows that in simple random sampling (without replacement) the estimator based on the sample mean is more efficient than Das' as well as Des Raj's estimators.

An unbiased estimate of Y^2 based on $u_r(s)$ and y_k ($k < r$), is given by

$$v_{rk}(s) = u_r(s)y_r + (N-1) u_r(s)y_k \quad (k < r = 1, 2, \dots, n). \quad \dots (5.3)$$

A uniformly better estimator than this is given by

$$\begin{aligned} E[v_{rk}(s)|T] &= \sum_{i=1}^n y_{(i)}^2 \frac{\left[\Sigma' \frac{1}{(N-1)} \dots \frac{1}{(N-r+1)} \frac{P(T|x_1, \dots, x_{r-1}, x_r = x_{(i)})}{P(T)} \right]}{P(T)} \\ &+ (N-1) \sum_{i \neq j=1}^n y_{(i)} y_{(j)} \frac{\left[\Sigma' \frac{1}{(N-1)} \dots \frac{1}{(N-r+1)} \left\{ P(T|x_1, \dots, x_k = x_{(i)}, \dots, \right. \right. \right. \\ &\quad \left. \left. \left. \frac{x_{r-1}, x_r = x_{(j)}) + P(T|x_1, \dots, x_k = x_{(j)}, \dots, x_{r-1}, x_r = x_{(i)})}{P(T)} \right\} \right]}{P(T)}. \quad \dots (5.4) \end{aligned}$$

This expression will also be identical for all r and k if and only if the sample is drawn by simple random sampling (without replacement).

Further, it may be seen on similar lines that in a more general (without replacement) sampling scheme which has been considered by Des Raj (1956), the estimators of Y (or of Y^2) obtained by improving Das' estimators will be identical if and only if the first unit in the sample is selected with pre-assigned probabilities and the remaining units are selected by simple random sampling (without replacement). For further reference about this, one may refer to Des Raj (1956) and Murthy (1957).

ACKNOWLEDGEMENT

I am grateful to Dr. D. Basu for his kind help and guidance in the preparation of the paper.

REFERENCES

- BASU, D. (1958): On sampling with and without replacement. *Sankhyā*, **20**, 287-294.
 DAS, A. C. (1951): Two-phase sampling and sampling with varying probabilities. *Bull. Int. Stat. Inst.*, **33**, 105-112.
 DES RAJ (1956): Some estimators in sampling with varying probabilities without replacement. *J. Amer. Stat. Ass.*, **51**, 274, 269-284.
 MURTHY, M. N. (1957): Ordered and unordered estimators in sampling without replacement. *Sankhyā*, **18**, 379-390.

Paper received: December, 1960.

Revised: April, 1961.

ON THE EVALUATION OF MOMENTS OF DISTINCT UNITS IN A SAMPLE

By P. K. PATHAK

Indian Statistical Institute

SUMMARY. In this paper, exact expressions for the moments of distinct units that appear in a sample, are derived under any sampling scheme. The importance of such a problem arises, e.g., when we select a simple random sample (with replacement) from a finite population and require the variance of the average of distinct units selected. It has been shown by Basu (1958) that this average is a better estimator of the population mean than the usual overall average.

1. PRELIMINARIES

In this section we give a lemma which will be used in the next section.

Lemma : *The coefficient $C_m(n)$ of*

$$Z_1^{\alpha_1} Z_2^{\alpha_2} \dots Z_m^{\alpha_m} \quad (\text{where } m \leq N, \alpha_i's > 0 \text{ and } \sum_{i=1}^m \alpha_i = n)$$

in the expansion of $(Z_1 + Z_2 + \dots + Z_N)^n$,

$$\text{is given by*} \quad C_m(n) = m^n - \binom{m}{1} (m-1)^n + \dots + (-1)^{m-1} \binom{m}{m-1} 1^n. \quad \dots (1.1)$$

In terms of the 'differences of zeros', $C_m(n)$ can be represented as

$$C_m(n) = \Delta^m O^n = \Delta^m x^n | x = 0, \quad \dots (1.2)$$

where Δ is the difference operator with unit increments. We shall be using freely these two expressions (1.1) and (1.2) for $C_m(n)$, whichever will be convenient to us in subsequent sections.

Corollary 1.1 : Coefficient of $Z_{i_1}^{\alpha_1} Z_{i_2}^{\alpha_2} \dots Z_{i_m}^{\alpha_m}$ (where $i_1, i_2, \dots, i_m$ are any m different integers chosen out of $1, 2, \dots, N$; $\alpha_i's > 0$ and $\sum_{i=1}^m \alpha_i = n$) in the expansion of $(Z_1 + Z_2 + \dots + Z_N)^n$, is given by $C_m(n)$.

Corollary 1.2 :

$$C_m(n) = m [C_m(n-1) + C_{m-1}(n-1)]. \quad \dots (1.3)$$

*Note that $C_m(n) = 0$ for $m > n$.

Corollary 1.3 : For all positive integral values of n and N , we have

$$N^n = \sum_{m=1}^N C_m(n) \binom{N}{m} \quad \dots (1.4)$$

2. MOMENTS OF DISTINCT UNITS

Positive order. Consider a population containing N units and a sampling scheme S for which

p_i = probability of inclusion of the i -th unit in the sample; ($i = 1, \dots, N$)

q_i = probability of exclusion of the i -th unit from the sample;

p_{ij} = probability of inclusion of the i -th and j -th units in the sample;

q_{ij} = probability of exclusion of the i -th and j -th units from the sample, etc.

We shall denote by v , the number of distinct units that appear in a sample. It is obvious that

$$v = Z_1 + Z_2 + \dots + Z_N, \quad \dots (2.1)$$

where $Z_i = \begin{cases} 1 & \text{if the } i\text{-th unit is included in the sample;} \\ 0 & \text{otherwise.} \end{cases}$

Now, by definition, if n is any positive integer, the n -th order moment of v is given by

$$\begin{aligned} E(v^n) &= E(Z_1 + Z_2 + \dots + Z_N)^n \\ &= E \left[\sum_{m=1}^N \Sigma_1 \Sigma_2 Z_1^{\alpha_1} \dots Z_m^{\alpha_m} \right], \quad \dots (2.2) \end{aligned}$$

where Σ_1 denotes the summation over $\binom{N}{m}$ combinations of mZ 's chosen out of $Z_1, Z_2, \dots, Z_N$ and Σ_2 denotes the summation over all products of the type

$$Z_1^{\alpha_1} Z_2^{\alpha_2} \dots Z_m^{\alpha_m} \quad (\alpha_i \text{'s} > 0; \sum_{i=1}^m \alpha_i = n).$$

Obviously, $E(\Sigma_2 Z_1^{\alpha_1} Z_2^{\alpha_2} \dots Z_m^{\alpha_m}) = p_{12 \dots m} C_m(n)$,

and therefore, $E(v^n) = \sum_{m=1}^N C_m(n) \Sigma_1 p_{12 \dots m}, \quad \dots (2.3)$

and

$$E(v^n) = \sum_{m=1}^N \binom{N}{m} C_m(n) p_{12 \dots m}$$

when $p_{i_1, i_2, \dots, i_m} = p_{12 \dots m}$ for every set of m distinct units $i_1, i_2, \dots, i_m$.

EVALUATION OF MOMENTS OF DISTINCT UNITS IN A SAMPLE

Negative order. To derive the negative moments of v of any order under any sampling scheme, we assume that $v \geq 1^*$, i.e.,

$$q_{12} \dots_N = 0$$

and define $u_i = 1 - Z_i$ ($i = 1, 2, \dots, N$).

Therefore, if t is any positive integer, the negative moment of v of order t is given by

$$E\left(\frac{1}{v^t}\right) = E\left[\frac{1}{(N - u_1 - u_2 - \dots - u_N)^t}\right] = \frac{1}{N^t} E\left[1 - \frac{\sum u_i}{N}\right]^{-t} \dots (2.4)$$

Since $0 < \sum_{i=1}^N u_i \leq (N-1)$, the infinite expansion is possible. Now let

$$(1-x)^{-t} = 1 + \sum_{r=1}^{\infty} A_r x^r,$$

so that $E\left(\frac{1}{v^t}\right) = \frac{1}{N^t} E\left[1 + \sum_{r=1}^{\infty} \frac{A_r}{N^r} (u_1 + u_2 + \dots + u_N)^r\right]. \dots (2.5)$

Since the infinite series $1 + \sum_{r=1}^{\infty} \frac{A_r}{N^r} (u_1 + u_2 + \dots + u_N)^r$ is bounded above by the absolutely convergent series

$$1 + \sum_{r=1}^{\infty} A_r \left(\frac{N-1}{N}\right)^r,$$

it, therefore, follows that

$$E\left(\frac{1}{v^t}\right) = \frac{1}{N^t} \left[1 + \sum_{r=1}^{\infty} \frac{A_r}{N^r} E(u_1 + \dots + u_N)^r\right]. \dots (2.6)$$

But it is apparent from (2.3) that

$$E(u_1 + u_2 + \dots + u_N)^r = E\left[\sum_{m=1}^N \Sigma_1 \Sigma_2 u_1^{\alpha_1} \dots u_m^{\alpha_m}\right]$$

$$= \sum_{m=1}^{(N-1)} (\Sigma_1 q_{12} \dots_m) \Delta^m x^r | x = 0.$$

*It is evident that this assumption is indeed necessary, otherwise no negative moment of v exists. In this paper, we restrict ourselves to those sampling schemes for which $P[v \geq 1] = 1$.

The N -th term vanishes by assumption $q_{12} \dots_N = 0$. Putting this in (2.6), we obtain

$$\begin{aligned} E \left(\frac{1}{v^t} \right) &= \frac{1}{N^t} \left[1 + \sum_{r=1}^{\infty} \frac{A_r}{N^r} \sum_{m=1}^{(N-1)} (\Sigma_1 q_{12} \dots_m) \Delta^m x^r \Big|_{x=0} \right] \\ &= \frac{1}{N^t} \left[1 + \sum_{m=1}^{(N-1)} (\Sigma_1 q_{12} \dots_m) \Delta^m \sum_{r=1}^{\infty} \frac{A_r}{N^r} x^r \Big|_{x=0} \right] \\ &= \frac{1}{N^t} \left[1 + \sum_{m=1}^{(N-1)} (\Sigma_1 q_{12} \dots_m) \Delta^m \left(1 - \frac{x}{N} \right)^{-t} \Big|_{x=0} \right], \end{aligned}$$

which on expansion gives

$$E \left(\frac{1}{v} \right) = \frac{1}{N^t} + \sum_{m=1}^{(N-1)} (\Sigma_1 q_{12} \dots_m) \left[\frac{1}{(N-m)^t} - \frac{\binom{m}{1}}{(N-m+1)^t} + \dots + (-)^m \frac{\binom{m}{m}}{N^t} \right]. \quad \dots (2.7)$$

In case, $q_{i_1, i_2, \dots, i_m} = q_{12} \dots_m$ for every set $(i_1, i_2, \dots, i_m)$ of m distinct units, this reduces to

$$= \frac{1}{N^t} + \sum_{m=1}^{(N-1)} \binom{N}{m} q_{12} \dots_m \left[\frac{1}{(N-m)^t} - \frac{\binom{m}{1}}{(N-m+1)^t} + \dots + (-)^m \frac{\binom{m}{m}}{N^t} \right].$$

Corollary 2.1: Putting $t = 1$ in the above result, we get

$$E \left(\frac{1}{v} \right) = \left[\frac{1}{N} + \sum_{m=1}^{(N-1)} (\Sigma_1 q_{12} \dots_m) \left\{ \frac{1}{(N-m)} - \frac{\binom{m}{1}}{(N-m+1)} + \dots + (-)^m \frac{\binom{m}{m}}{N} \right\} \right].$$

Since
$$\frac{1}{(N-m)} - \frac{\binom{m}{1}}{(N-m+1)} + \dots + (-)^m \frac{\binom{m}{m}}{N} = \frac{m!}{N(N-1) \dots (N-m)},$$

$$\therefore E \left(\frac{1}{v} \right) = \frac{1}{N} \left[1 + \frac{1}{(N-1)} \Sigma_1 q_1 + \frac{1.2}{(N-1)(N-2)} \Sigma_1 q_{12} + \dots + \frac{1.2 \dots (N-1)}{(N-1) \dots 2.1} \Sigma_1 q_{12} \dots_{N-1} \right]. \quad \dots (2.8)$$

In case, $q_{i_1, i_2, \dots, i_m} = q_{12} \dots_m$ for every set $(i_1, i_2, \dots, i_m)$ of m distinct units, (2.8) reduces to

$$E \left(\frac{1}{v} \right) = \frac{1}{N} + \sum_{m=1}^{(N-1)} \frac{q_{12} \dots_m}{(N-m)} \quad \dots (2.8a)$$

Particular cases: (a) *Simple random sampling (with replacement).*
In this sampling scheme

$$q_{12} \dots_m = \frac{(N-m)^n}{N^n},$$

where n is the size of the sample.

EVALUATION OF MOMENTS OF DISTINCT UNITS IN A SAMPLE

$$\therefore E\left(\frac{1}{v}\right) = \frac{1}{N} + \sum_{m=1}^{(N-1)} \frac{(N-m)^n}{N^n(N-m)} = \frac{1^{n-1} + 2^{n-1} + \dots + N^{n-1}}{N^n} \dots (2.8b)$$

In terms of Bernoulli numbers, (2.8b) is given by (Davis, 1935, p. 188)

$$E\left(\frac{1}{v}\right) = \frac{1}{n} + \frac{1}{2N} + \frac{1}{n} \sum_{s=1}^{\leq \frac{n-1}{2}} (-)^{s-1} \binom{n}{2s} \frac{B_s}{N^{2s}}, \dots (2.9)$$

where B_s is the s -th Bernoulli number.

For large N , this gives us a very convenient method for computing $E\left(\frac{1}{v}\right)$.

(b) *Simple random sampling (without replacement).*

Here,

$$q_{12} \dots m = \begin{cases} \frac{\binom{N-m}{n}}{\binom{N}{n}} & \text{for } m \leq N-n; \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} \therefore E\left(\frac{1}{v}\right) &= \left[\frac{1}{N} + \frac{\binom{N-1}{n}}{(N-1)\binom{N}{n}} + \dots + \frac{\binom{n}{n}}{n\binom{N}{n}} \right] \\ &= \frac{1}{\binom{N}{n}} \left[\frac{1}{n} + 1 + \frac{(n+1)}{1.2} + \frac{(n+1)(n+2)}{1.2.3} \right. \\ &\quad \left. + \dots + \frac{(n+1)(n+2)\dots(N-2)(N-1)}{1.2.3 \dots (N-n)} \right]. \end{aligned}$$

Combining the terms one by one, we get

$$E\left(\frac{1}{v}\right) = \frac{1}{\binom{N}{n}} \frac{(n+1)(n+2)\dots N}{n \cdot 1.2.3 \dots (N-n)} = \frac{1}{n},$$

which is in agreement with the process of sampling.

Corollary 2.2: For any integer t ($\neq 0$), it can be shown in a similar manner that

$$\begin{aligned} E(v^t) &= N^t + \sum_{m=1}^{(N-1)} (\Sigma_1 q_{12} \dots m) \Delta^m (N-x)^t \Big|_{x=0} \\ &= N^t + \sum_{m=1}^{N-1} (\Sigma_1 q_{12} \dots m) [(N-m)^t - \binom{m}{1}(N-m+1)^t \\ &\quad + \dots + (-)^m \binom{m}{m} N^t]. \dots (2.10) \end{aligned}$$

The following theorem giving the expected values of certain functions of v , is an obvious generalization of (2.10).

Theorem : If $f(Z)$ is any function of Z defined in the domain, $0 < Z \leq N$, for which the infinite expansion in powers of Z is possible, and if the expectation can be taken term by term, then

$$E[f(v)] = f(N) + \sum_{m=1}^{(N-1)} (\sum_1 q_{12} \dots q_m) [f(N-m) - \binom{m}{1} f(N-m+1) + \dots + (-)^m f(N)]. \quad \dots \quad (2.11)$$

Proof: Express $f(Z)$ in the form

$$f(Z) = \sum_{r=-\infty}^{\infty} A_r Z^r.$$

By assumption
$$E[f(v)] = E \left[\sum_{r=-\infty}^{\infty} A_r v^r \right] = \sum_{r=-\infty}^{\infty} A_r E(v^r).$$

Putting the value of $E(v^r)$ from (2.10), we get

$$\begin{aligned} E[f(v)] &= \sum_{r=-\infty}^{\infty} \left\{ A_r \left[N^r + \sum_{m=1}^{(N-1)} (\sum_1 q_{12} \dots q_m) \Delta^m (N-x)^r \Big|_{x=0} \right] \right\} \\ &= f(N) + \sum_{m=1}^{(N-1)} (\sum_1 q_{12} \dots q_m) \Delta^m f(N-x) \Big|_{x=0}, \end{aligned}$$

which on expansion gives (2.11).

ACKNOWLEDGEMENT

I am indebted to Dr. D. Basu for his keen interest and generous guidance in the preparation of the paper.

REFERENCES

- BASU, D. (1958) : On sampling with and without replacement. *Sankhyā*, **20**, 287-294.
DAVIS, H. T. (1935) : *Tables of the Higher Mathematical Function*, Vol. II, The Principia Press, Inc., Bloomington, Indiana.

Paper received : October, 1960.

MOMENTS OF FIRST ORDER AUTOCORRELATION OF THE SECOND ORDER AUTOREGRESSIVE SCHEME

By SUKUMAR MUKHERJEE

Patna University, India

SUMMARY. The paper considers a realization $\{x_t\}$ of a time series obeying Yule Scheme, $x_{t+2} - \alpha_1 x_{t+1} - \alpha_2 x_t = \varepsilon_{t+2}$ where the ε 's have zero means and variance σ^2 , independent of t , α_1 and α_2 being, of course, constants. Under the hypothesis that $\alpha_2 \neq 0$, expressions for the first and second raw moments of the first order autocorrelation (circularity of autocovariances is assumed) has been given.

1. INTRODUCTION

Let $\{x_i\}$ ($i = 1, 2, \dots, n$) be a realization of a time-series, with theoretical mean of the series zero and the underlying model

$$x_{t+2} - \alpha_1 x_{t+1} - \alpha_2 x_t = \varepsilon_{t+2}$$

where ε 's follow normal distribution with mean zero and variance σ^2 , α_1 and α_2 are constants. Giving circular definition to autocovariances the joint distribution of r_1 and r_2 , uncorrected for mean, can be written in the form (Quenouille, 1949),

$$p(r_1 r_2 / \alpha_1 \alpha_2) = \frac{\Gamma(\frac{1}{2}n+1) \Gamma(\frac{1}{2}n)}{\Gamma(\frac{n+1}{2}) \Gamma(\frac{n-1}{2})} \frac{1}{\pi} Q_2^{\frac{1}{2}(n-3)} (A + 2B_j r_j)^{-\frac{n}{2}} \dots \quad (1.1)$$

where

$$Q_2 = R_2/R_1 = \begin{vmatrix} 1 & r_1 & r_2 \\ r_1 & 1 & r_1 \\ r_2 & r_1 & 1 \end{vmatrix} \div \begin{vmatrix} 1 & r_1 \\ r_1 & 1 \end{vmatrix}$$

$$A = 1 + \alpha_1^2 + \alpha_2^2$$

$$B_j r_j = -\alpha_1(1 - \alpha_2)r_1 - \alpha_2 r_2.$$

Jenkins (1954) has shown that the form (1.1) gives correct marginal moments for r_1 but not for r_2 and suggested the distribution given in (1.3). Writing (1.1) in the form

$$p(r_1 r_2 / \alpha_1 \alpha_2) = p(r_1 r_2) \{1 + \alpha_1^2 + \alpha_2^2 - 2\alpha_1(1 - \alpha_2)r_1 - 2\alpha_2 r_2\}^{-\frac{n}{2}} \dots \quad (1.2)$$

where $p(r_1 r_2)$ is the distribution when $\alpha_1 = \alpha_2 = 0$, the transformations $r = r_1$ and

$$v = \frac{r_2 - r_1^2}{1 - r_1^2} \text{ give}$$

$$p(rv / \alpha_1 \alpha_2) = p(r) p(v) \{1 + \alpha_1^2 + \alpha_2^2 - 2\alpha_1(1 - \alpha_2)r - 2v\alpha_2(1 - r^2) - 2\alpha_2 r^2\}^{-\frac{n}{2}} \dots \quad (1.3)$$

where

$$p(r_1) = p(r_1 / \alpha_1 = 0 = \alpha_2), \quad p(v) = p(v / \alpha_1 = 0 = \alpha_2)$$

and v is independently distributed of r_1 with distribution

$$p(v) = k(1 - v^2)^{\frac{1}{2}(n-2)} (1 - v) \dots \quad (1.4)$$

Moments of (1.4) are given by

$$m'(2k) = \frac{(2k-1)(2k-3)\dots 1}{(n+1)(n+3)\dots(n+2k-1)}, \quad m'(2k-1) = -m'(2k).$$

In particular

$$E(v) = -\frac{1}{n+1}, \quad E(v^2) = \frac{1}{n+1}, \quad E(v^3) = -\frac{1.3}{(n+1)(n+3)}, \quad E(v^4) = \frac{1.3}{(n+1)(n+3)}.$$

Since the mean and variance are generally needed for usual test of significance only the first and second order moments of r about the origin are quoted below. Some of the relations given below are obtained after very lengthy algebraic reductions; here the details of the derivations are omitted.

2. SOME NOTATIONS AND RESULTS TO BE USED SUBSEQUENTLY

Let $A = 1 + \alpha_1^2 + \alpha_2^2 - 2\alpha_1(1 - \alpha_2)r - 2\alpha_2r^2$ and $B = 2\alpha_2(1 - r^2)v$, where r stands for r_1 ,

so
$$\left. \frac{dA}{dr} \right|_{r=0} = 2\alpha_1(1 - \alpha_2) = A'_0$$

and
$$\left. \frac{d^2A}{dr^2} \right|_{r=0} = -4\alpha_2 = A''_0 \quad \text{i.e., } \alpha_2 = -\left(\frac{A''_0}{4}\right),$$

also,
$$A_0 = 1 + \alpha_1^2 + \alpha_2^2.$$

If
$$A^{-t} = k_0 + k_1r + k_2r^2 + \dots$$

we have

$$k_0 = (1 + \alpha_1^2 + \alpha_2^2)^{-t} = A_0^{-t}$$

$$k_1 = -tA_0^{-t-1} A'_0$$

$$2!k_2 = t(t+1)A_0^{-t-2} (A'_0)^2 - tA_0^{-t-1} A''_0$$

$$3!k_3 = -t(t+1)(t+2)A_0^{-t-3} (A'_0)^3 + 3t(t+1)A'_0 A''_0 A_0^{-t-2}$$

$$4!k_4 = t(t+1)(t+2)(t+3)A_0^{-t-4} (A'_0)^4 - 6t(t+1)(t+2)A_0^{-t-3} (A'_0)^2 A''_0 + 3t(t+1)A_0^{-t-2} (A''_0)^2 \text{ etc.}$$

If $t = \frac{n}{2}$ we have $p_{2t}(r)$ = the density function of r based on n observations calculated from a sample of independently distributed normal variables, so

$$p_{2t}(r) = \frac{\Gamma\left(\frac{n}{2} + 1\right)}{\Gamma\left(\frac{1}{2}\right) \Gamma\left(\frac{n+1}{2}\right)} (1-r^2)^{\frac{n-1}{2}} \quad (\text{Dixon, 1944}),$$

therefore

$$\begin{aligned} (1-r^2)p_{2t}(r) &= \frac{\Gamma\left(\frac{n}{2} + 1\right)}{\Gamma\left(\frac{1}{2}\right) \Gamma\left(\frac{n+1}{2}\right)} (1-r^2)^{\frac{n}{2} + \frac{1}{2}} = \frac{\left(\frac{n+1}{2}\right) \Gamma\left(\frac{n+2}{2} + 1\right)}{\left(\frac{n}{2} + 1\right) \Gamma\left(\frac{1}{2}\right) \Gamma\left(\frac{n+3}{2}\right)} (1-r^2)^{\frac{n+2}{2} - \frac{1}{2}} \\ &= \frac{2t+1}{2t+2} p_{2(t+1)}(r). \end{aligned}$$

ON AUTOREGRESSIVE SCHEME

3. FIRST MOMENT OF r

$$\begin{aligned} p(rv) &= p(r)p(v)\{A-Bv\}^{-\frac{n}{2}} = p(r)p(v)\{A-Bv\}^{-t} \\ &= p(r)A^{-t}\left\{1+t\frac{B}{A}v+\frac{t(t+1)}{2!}\left(\frac{B}{A}\right)^2v^2+\dots\right\}p(v) \end{aligned}$$

so

$$\begin{aligned} p(r) &= \int_{-1}^1 p(rv)dv = A^{-t}p_{2t}(r) + \frac{2t}{2t+2}\left(\frac{A_0''}{4}\right)p_{2(t+1)}(r)A^{-t-1} \\ &\quad + \frac{2t(2t+3)}{2\cdot(2t+4)}\left(\frac{A_0''}{4}\right)^2A^{-t-2}p_{2(t+2)}(r)+\dots \dots \dots (3.1) \end{aligned}$$

Now

$$\int_{-1}^1 A^{-t}p_{2t}(r)dr = \left\{k_0 + \frac{k_2}{2!}\frac{1}{2t+2} + \frac{k_4}{4!(2t+2)(2t+4)} + \dots\right\}$$

and

$$\begin{aligned} \int_{-1}^1 rA^{-t}p_{2t}(r)dr &= \frac{-tA_0'}{2t+2}\int_{-1}^1 A^{-t-1}p_{2(t+1)}(r) - \frac{t}{2t+2}A_0''\int_{-1}^1 rA^{-t-1}p_{2(t+1)}(r)dr \\ &= \frac{-tA_0'}{2t+2}\left\{\int_{-1}^1 A^{-t-1}p_{2(t+1)}(r)dr - \frac{t+1}{2t+4}A_0''\int_{-1}^1 A^{-t-2}p_{2(t+2)}(r)dr + \dots\right\} \end{aligned}$$

using this result we have from (3.1) after some algebraic reduction

$$\begin{aligned} m'(1) &= \int_{-1}^1 rp(r)dr = \frac{-tA_0'}{2t+2}\left\{1 - \frac{4(t+1)}{(2t+4)}\left(\frac{A_0''}{4}\right) + \frac{6(t+1)}{(2t+6)}\left(\frac{A_0''}{4}\right)^2 \right. \\ &\quad \left. - \frac{8(t+1)}{(2t+8)}\left(\frac{A_0''}{4}\right)^3 + \frac{10(t+1)}{(2t+10)}\left(\frac{A_0''}{4}\right)^4 - \dots\right\} \dots (3.2) \end{aligned}$$

Now in order that the generated series be of finite values for large n , $|\alpha_2|$ should be less than 1. Hence for finite n the series on the right hand side of (3.2) will converge.

4. SECOND MOMENT OF r

From (3.1) we have

$$\begin{aligned} (1-r^2)p(r) &= \frac{2t+1}{2t+2}A^{-t}p_{2(t+1)}(r) + \frac{2t(2t+3)}{(2t+2)(2t+4)}\left(\frac{A_0''}{4}\right)A^{-t-1}p_{2(t+2)}(r) + \dots \\ &\dots (4.1) \end{aligned}$$

Since $\frac{\partial}{\partial \alpha_1} A^{-t} = -t A^{-t-1} [2\alpha_1 - 2(1-\alpha_2)r]$ we have from (4.1)

$$\begin{aligned} \frac{\partial}{\partial \alpha_1} (1-r^2)p(r) = & -[\alpha_1 - (1-\alpha_2)r](2t) \left\{ \frac{2t+1}{2t+2} p_{b(t+1)}(r) + \frac{2}{2t+4} \left(\frac{A_0''}{4}\right) p_{b(t+2)}(r) \right. \\ & \left. + \frac{2t+3}{2t+6} \left(\frac{A_0''}{4}\right)^2 p_{b(t+3)}(r) + \dots \right\} \quad \dots (4.2) \end{aligned}$$

where $p_{b(t+s)}(r)$ = the density function of r based on $2(t+s)$ observations. That is the density function given in (3.1) where t is replaced by $t+s$. After some algebraic reduction we have

$$\begin{aligned} \frac{\partial}{\partial \alpha_1} \int_{-1}^1 (1-r^2) p(r) dr = & -2t\alpha_1 \left\{ \frac{2t+1}{2t+2} + \frac{2}{2t+4} \left(\frac{A_0''}{4}\right) + \frac{2t+3}{2t+6} \left(\frac{A_0''}{4}\right)^2 + \frac{4}{2t+8} \left(\frac{A_0''}{4}\right)^3 + \dots \right\} \\ & + 2t(1-\alpha_2) \left\{ \frac{2t+1}{2t+2} m'_{b(t+1)}(1) + \frac{2}{2t+4} \left(\frac{A_0''}{4}\right) m'_{b(t+2)}(1) + \frac{2t+3}{2t+6} \left(\frac{A_0''}{4}\right)^2 m'_{b(t+3)}(1) + \dots \right\} \quad \dots (4.3) \end{aligned}$$

where $m'_{b(t+s)}(1)$ = the first moment of r based on $2(t+s)$ observations. That is the expression obtained from (3.2) when t is replaced by $t+s$ whence

$$\begin{aligned} \int_{-1}^1 r^2 p(r) dr = m'(2) = & 2t \left\{ \frac{2t+1}{2t+2} + \frac{2}{2t+4} \left(\frac{A_0''}{4}\right) + \frac{2t+3}{2t+6} \left(\frac{A_0''}{4}\right)^2 + \dots \right\} \frac{\alpha_1^2}{2} \\ & + \frac{2t(1-\alpha_2)}{2} \left\{ \frac{2t+1}{2t+4} \left[1 - \frac{4(t+2)}{2t+6} \left(\frac{A_0''}{4}\right) + \frac{6(t+2)}{2t+8} \left(\frac{A_0''}{4}\right)^2 - \dots \right] \right. \\ & + \frac{2}{2t+6} \left(\frac{A_0''}{4}\right) \left[1 - \frac{4(t+3)}{2t+8} \left(\frac{A_0''}{4}\right) + \frac{6(t+3)}{2t+10} \left(\frac{A_0''}{4}\right)^2 + \dots \right] \\ & + \frac{2t+3}{2t+8} \left(\frac{A_0''}{4}\right)^2 \left[1 - \frac{4(t+4)}{2t+10} \left(\frac{A_0''}{4}\right) + \frac{6(t+4)}{2t+12} \left(\frac{A_0''}{4}\right)^2 + \dots \right] \\ & \left. + \dots \right\} \int A_0' d\alpha_1 + C. \quad \dots (4.4) \end{aligned}$$

ON AUTOREGRESSIVE SCHEME

Where c is an arbitrary constant. From initial conditions we find $c = 1/2t+2$ as first approximation. Thus

$$m'(2) = \frac{1}{2t+2} + \alpha_1^2 t \left\{ \frac{2t+1}{2t+2} + \frac{2}{2t+4} \left(\frac{A_0''}{4} \right) + \frac{2t+3}{2t+6} \left(\frac{A_0''}{4} \right)^2 + \frac{4}{2t+8} \left(\frac{A_0''}{4} \right)^3 \right. \\ \left. + \frac{2t+5}{2t+10} \left(\frac{A_0''}{4} \right)^2 + \dots - (1-\alpha_2)^2 \left[\frac{2t+1}{2t+4} - \frac{4t}{2t+6} \left(\frac{A_0''}{4} \right) + \frac{2(4t+1)}{2t+8} \left(\frac{A_0''}{4} \right)^2 \right. \right. \\ \left. \left. - \frac{12t}{2t+10} \left(\frac{A_0''}{4} \right)^3 + \dots \right] \right\}. \quad \dots (4.5)$$

Particular cases. (i) If $\alpha_2 = 0$ and $\alpha_1 \neq 0$ we have from (3.2) and (4.5)

$$m'(1) = \frac{2\alpha_1 t}{2t+2} = \frac{n\alpha_1}{n+2} \quad \dots (4.6)$$

and

$$m'(2) = \frac{1}{2t+2} + \alpha_1^2 t \left\{ \frac{2t+1}{2t+2} - \frac{2t+1}{2t+4} \right\} \\ = \frac{1}{2t+2} \left\{ 1 + \frac{2t(2t+1)}{2t+4} \alpha_1^2 \right\} \\ = \frac{1}{n+2} \left\{ 1 + \frac{n(n+1)}{n+4} \alpha_1^2 \right\}. \quad \dots (4.7)$$

(ii) If $\alpha_1 = 0 = \alpha_2$ we have from (3.2) and (4.5)

$$m'(1) = 0 \quad \dots (4.8)$$

and

$$m'(2) = 1/n+2 \quad \dots (4.9)$$

Results (4.6) and (4.7) agree with the corresponding moments given by Leipnik (1947) and (1958), Jenkins (1956), Author (1958). Results (4.8) and (4.9) agree with the corresponding moments given by Dixon (1944).

I am greatly indebted to Dr. D. N. Lal of the Department of Statistics, Patna University for his advice and encouragement during the preparation of this paper.

REFERENCES

- DIXON, W. J. (1944): Further contributions to the problem of serial correlation. *Ann. Math. Stat.*, **14**, 119.
- JENKINS, G. M. (1954): Test of hypothesis in the linear autoregressive model, 1. *Biometrika*, **41**, 405.
- (1956): Test of hypothesis in the linear autoregressive model, 2. *Biometrika*, **43**, 186.
- Leipnik, R. B. (1947): Distribution of the serial correlation in a circularly correlated universe. *Ann. Math. Stat.*, **18**, 80.
- (1958): Note on the characteristic function of a serial-correlation distribution. *Biometrika*, **45**, 559.
- Quenouille, M. H. (1959): The joint distribution of serial correlation coefficients. *Ann. Math. Stat.*, **20**, 561.

Paper received : November, 1960.

CORRIGENDA

A Note On Mutually Orthogonal Latin Squares: By S. S. Shrikhande, *Sankhyā* 23, Series A, 115-116.

- (1) Summary, line 1: should read 'If $n \neq 4$, it is proved' instead of 'It is proved'.
- (2) Page 116, line 13: should read 'Adding to the blocks in (2.1) further $2n$ blocks' instead of 'Forming $2n$ block'.

ACKNOWLEDGEMENT

Thanks are due to the following for their kind help in the editing of the papers published in this issue.

Dr. D. Basu

Dr. C. R. Rao

Shri N. S. Iyengar

Dr. J. Roy

Shri K. R. Parthasarathy

Shri S. R. S. Varadhan

Editor,

Sankhyā : The Indian Journal of Statistics.



INDIAN STATISTICAL INSTITUTE

President

SHRI C. D. DESHMUKH

Chairman

SIR D. N. MITRA

Honorary Fellows

PROFESSOR CORRADO GINI
SIR R. A. FISHER, SC.D., F.R.S.

Vice-Presidents

DR. S. K. BANERJI, D.SC.
PROFESSOR S. N. BOSE, F.R.S.
PROFESSOR D. R. GADGIL, M.A., M.LITT.
DR. ZAKIR HUSAIN, PH.D.
DR. Q. M. HUSSAIN, M.A., PH.D.
PROFESSOR D. G. KARVE, M.A.
PROFESSOR K. B. MADHAVA, M.A., A.I.A.
SIR SHRI RAM

Vice-Chairmen

SHRI K. P. GOENKA
SHRI K. C. MAHINDRA, M.A. (Cantab)
SHRI D. N. MUKHERJEE, M.L.A.
SHRI S. C. RAY, M.A., B.L.
PROFESSOR N. R. SEN, M.A., D.SC.

Treasurer

DR. S. C. LAW, PH.D., F.N.I.

Secretary

PROFESSOR P. C. MAHALANOBIS, F.R.S.

Joint Secretaries

SHRI S. BASU, M.SC.
SHRI N. C. CHAKRAVARTI, M.A.
SHRI P. PANT, M.SC.

Branch Secretaries

SHRI V. M. DANDEKAR, M.A. (POONA BRANCH)
DR. D.V. RAJALAKSHMAN, (MADRAS BRANCH)
MR. A. SALIM, M.A., M.SC. (ALIGARH BRANCH)
DR. G. R. SETH, (DELHI BRANCH)
PROFESSOR SRINAGABHUSHANA,
(MYSORE STATE BRANCH)
SHRI M. A. TELANG, M.A. (BOMBAY BRANCH)

The Indian Statistical Institute is a Society registered in April 1932 under the Indian Societies Registration Act (Act No. XXI of 1860). Since 1 April 1960, it has been declared as an Institution of National Importance by the Indian Statistical Institute Act No. 57 of 1959.

Extracts from the Memorandum of Association

II. *The objects of the Institute shall be:—*

- (1) *To promote the study and dissemination of knowledge of and research on Statistics and other subjects relating to planning for national development and social welfare.*
- (2) *To provide for and undertake the collection of information, investigations, projects, and operational research for purposes of planning and the improvement of efficiency of management and production.*
- (3) *To undertake any other activities which in the opinion of the Council may be usefully carried out by the Institute in the furtherance of national development and social welfare.*

IV. *The membership of the Institute and its constituent bodies shall be open to all persons irrespective of sex, nationality, race, creed or class.*

V. *The Institute shall not distribute any part of its income or profit as dividend, gift, or bonus or in any other way among its members, provided nevertheless that office-bearers, workers and other employees shall not be debarred from receiving remuneration or participating in grants because of their also being members of the Institute.*

The Institute consists of Ordinary and Life Members, Institutional Members, Fellows, Honorary Members and Patrons. The supreme control including the power of making rules is vested in the members in General Meeting assembled. The Chairman and other office-bearers are elected once every two years. The management of the affairs of the Institute is in the hands of a Council elected annually by the members of the Institute.

Facilities are provided for student and sessional membership.

Membership fee, Rs. 25/- (foreign Rs. 30/-) per year, Institutional Membership fee, Rs. 100/- per year include subscription to one series of *Sankhyā: The Indian Journal of Statistics*.

Detailed information and Membership forms may be obtained on application to the Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta-35.

SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS

EDITED BY P. C. MAHALANOBIS, F.R.S.

Sankhyā is published in co-operation with the Indian Statistical Institute in two separate Series. *Series A* is for papers and notes on new methods and techniques. *Series B* is mainly for original data, records and applications of new methods. Each series of *Sankhyā* will contain about 400 pages per volume with plates and tables (to be issued in approximately four equal parts) which is to include communications from the Indian Statistical Institute. Papers intended to be treated as communications from the Institute should be addressed to the Journal Committee (203, Barrackpore Trunk Road, Calcutta-35) appointed by the Indian Statistical Institute, which consists of the following members:

Journal Committee

Professor S. N. Bose, *National Professor, (President).*
Shri G. Basu, *Federation of Indian Chambers of Commerce and Industry, Calcutta.*
Shri V. M. Dandekar, *Gokhale Institute of Politics and Economics, Poona.*
Shri D. B. Lahiri, *Indian Statistical Institute, Calcutta.*
Professor K. B. Madhava, *Madras.*
Professor P. C. Mahalanobis, *(Editor of Sankhyā, Ex-Officio).*
Dr. Sujit Kumar Mitra, *Indian Statistical Institute.*
Shri M. Mukherjee, *Indian Statistical Institute, Calcutta.*
Dr. K. R. Nair, *Central Statistical Organisation, New Delhi.*
Shri Pitambar Pant, *Indian Statistical Institute.*
Dr. B. N. Prasad, *Allahabad University.*
Dr. C. R. Rao, *Indian Statistical Institute, Calcutta.*
Dr. J. Roy, *Indian Statistical Institute.*
Shri A. Mahalanobis, *Indian Statistical Institute, (Editorial Secretary).*

All other papers for publication and books and reprints for review should be sent to the Editor, *Sankhyā*, 203, Barrackpore Trunk Road, Calcutta-35.

Every member of the Indian Statistical Institute will get either of the two series of the Journal free of cost. The subscription price per volume including packing and postage for each of the series for non-members, payable in advance, is: USA and Canada: \$ 10.00; Other Countries: £ 3; India: Rs. 30.00. The subscription price per volume for both the series taken together is: USA and Canada: \$ 18.00; Other Countries: £ 5-6s; India: Rs. 50.00. Subscriptions should be sent to the Secretary, Statistical Publishing Society, 204/1, Barrackpore Trunk Road, Calcutta-35.

It is a condition of publication in *Sankhyā* that the paper shall not already have been published elsewhere and will not be reprinted without leave of the Editor. Authors are requested to send two typewritten copies of each manuscript. It is necessary that the raw data or measurements on which the paper is based, and detailed information of their sources, should be furnished in every case. It is desirable that the distribution or correlation charts from which statistical constants are calculated should accompany the paper. Tables should be numbered consecutively with Arabic figures without brackets (e.g. Table 1, Table 45 etc.) and should have short explanatory headings. Diagrams and drawings should also be numbered consecutively with Arabic figures and should be sent in a finished state suitable for direct reproduction with all lettering in pencil. References to tables and diagrams should be given within brackets in the text. The list of literature cited should be arranged alphabetically according to authors' names. In the case of journals the volume, number, year and page references, and in the case of printed books the name of the author and the publisher, the date of publication, and number of the edition or impression should always be stated clearly. Abbreviations, local names, vernacular measures, etc., should be written very legibly (preferably in block letters) and should be explained where they first occur in the text.

Contributors receive 25 copies (joint authors: 15 copies each) of their papers free. The cost of additional reprints will depend upon the number required; for every 50 additional copies the rate is Rs. 12/- for each sheet of eight pages or Rs. 2/- per page for part of a sheet, with an extra charge for plates printed on special paper. Additional copies should be ordered at the time of returning the final proof.

Publication of papers, notes, reviews, discussion, etc. does not signify the Editor's approval of the arguments used or conclusions drawn by the authors. It is to be distinctly understood that all opinions expressed are purely individual and no responsibility for such opinions rests on the Editor, the Journal Committee or the Indian Statistical Institute.

Printed by Kalipada Mukherjee at EKA PRESS, 204/1 Barrackpore Trunk Road, Calcutta-35 and published by Prafulla Chandra Mahalanobis from Statistical Publishing Society, 204/1 B. T. Road, Calcutta-35



